# The Food Scene around Chapman University

Spencer Au, Miles Allen, Daniel Boudagian

## Introduction

**Variables:**

The variables we used were a mix of categorical (such as red, blue, white, etc) and continuous (such as numerical values like 1 or 2). The categorical values were RestaurantType, Culture, Specialty, AverageMealPrice (ranging from $ to $$$$), HealthInspectionRating, and AlcoholAvailability. The continuous values were YearOpened, DistancefromChapman, Rating, Reviews, CompetitorDensity (which is the number of similar restaurants in the general vicinity), YearsSinceOpen, and Competitiveness (which we got from the Rating / (CompetitorDensity + 1)

**Data:**

The data we are using is a dataset of 147 restaurants around Chapman University, in which we had to scrape the values manually through Yelp and Google Maps information. There are obviously some issues with the data, such as the fact that there is not a lot of variance for a lot of the categories, such as Culture and Specialty. For example, within the City of Orange, there aren't really a lot of Cuban culture restaurants or places that specialize in a tasting menu versus say a restaurant that has an American culture or specializes in cafe food such as coffee, tea, and pastries. That being said, we did try to include 4 or 5 restaurants for each culture and specialty respectively, and attempted to generalize each category when possible. As a result, we had to differ from some of the original analysis plan questions and methods as either the initial questions or methods were not feasible given this specific dataset.

## Q1: (Supervised Model) Impact of Culture, Specialty, Price and Alcohol on Longevity: How do the cultural background of the cuisine, its specialty, and the availability of alcohol influence a restaurant's operational lifespan?

### Methods:

We are using Restaurant Type (Categorical), Culture (Categorical), Specialty (Categorical), Alcohol Availability (Categorical), Price (Categorative) to predict the Years in Operation (Continuous) a restaurant will be. We perform Cleaning by handling missing values appropriately for each variable and use OneHotEncoder() for Categorical Values. Basically, we clean our data by dropping any missing information and using a technique called "OneHotEncoder" to neatly organize different categories like restaurant type and alcohol availability. We will use a technique called a Gradient Boosting Tree to study how factors like cultural background, specialty, price, and alcohol availability affect how long a restaurant stays in business. To ensure accuracy, we'll divide our data, using 80% for training our model and 20% for testing it. The model will then predict the operational lifespan of restaurants based on these factors.

### Results:

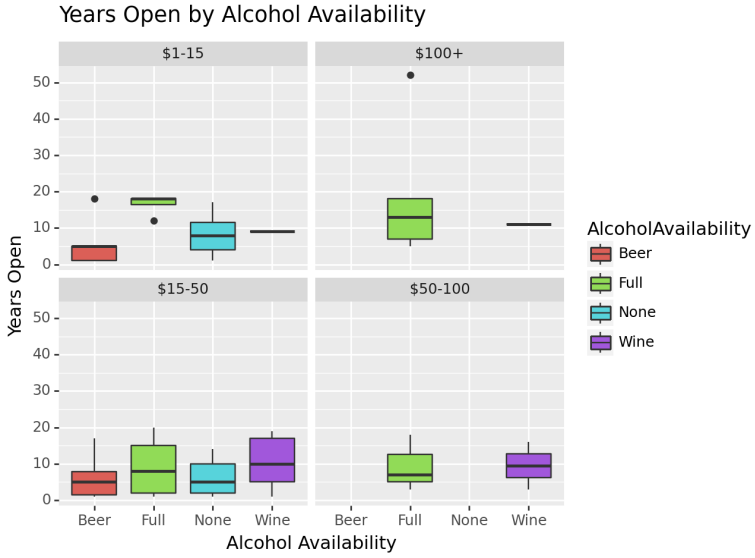
| | Training Set | Testing Set |
|---|---|---|
| MSE | 26.2273 | 23.2827 |
| MAE | 3.4085 | 3.8396 |
| MAPE | 0.9727 | 1.1273 |
| R2 | 0.4534 | 0.3584 |

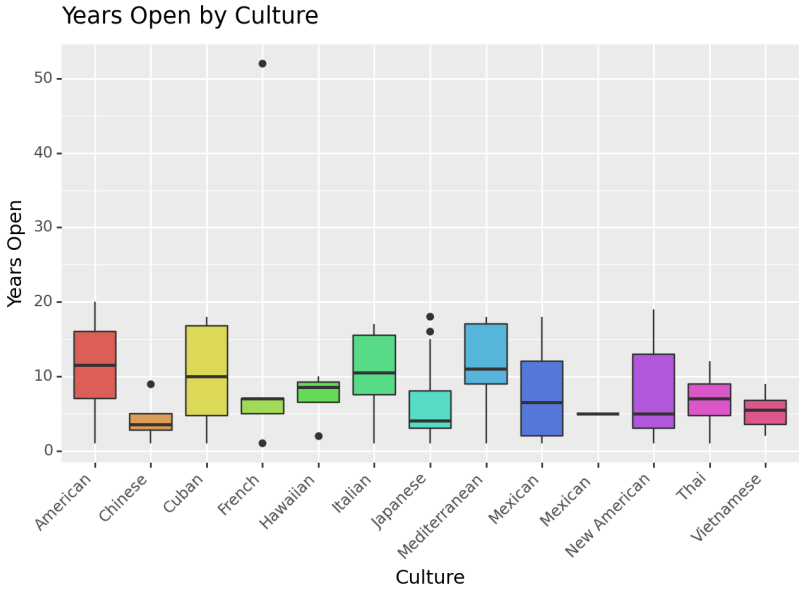
We decided to go with a Gradient Boosting Tree instead of the alternatives such as a Linear Regression Model (with and without polynomial features), as well the Lasso and Ridge Models and the other two tree based models we have at our disposal. This decision was made based on its superior performance. Specifically, the Gradient Boosting Tree showed better consistency and accuracy in predicting how long restaurants stay open. Its scores in measures like MSE (Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R2 (a measure of how well the model fits the data) were more reliable and closer to real-world values, both when training the model with our data and when testing its predictions, compared to the other models we considered.
In addition, the GBT also had a much lower rate of overfitting as it performed relatively consistently across both the training and testing sets, versus the other models where they each seemed to express a high degree of overfitting, where they performed well on the training set

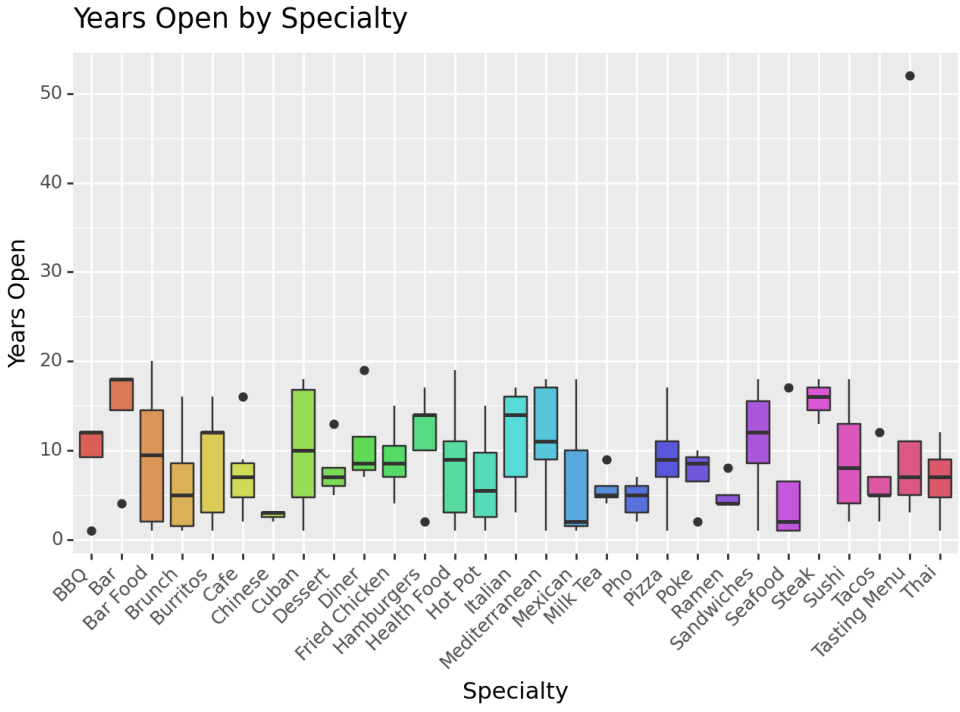
but rather terribly on the testing set. This indicates that they might not generalize well to new, unseen data.

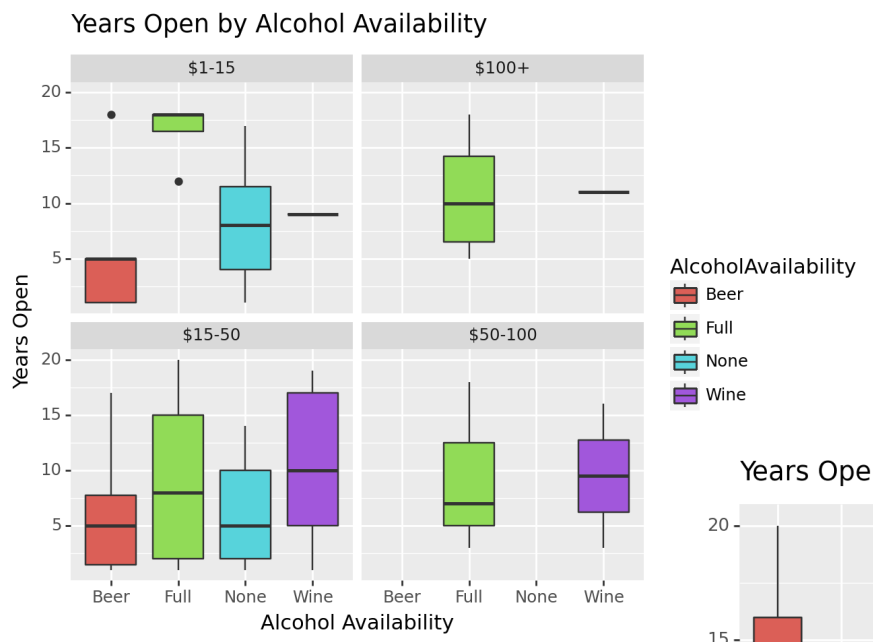
**Graphs:**

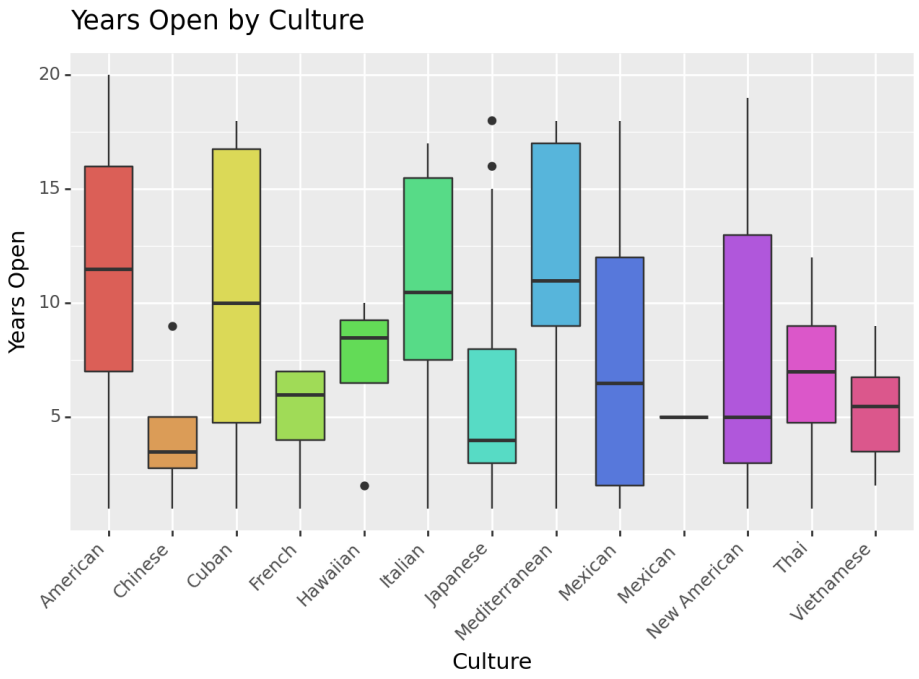
- P1: Boxplot with faceted graphs for different price ranges of $1-15, $15-50, $50-100, and $100+. The graph plots the years that restaurants have been open for each of the alcohol availability categories
- P2: Boxplot that plots the years since open for different cultures
- P3: Boxplot that plots the years since open for different specialties
- Boxplots for P1, P2, and P3 that exclude the one outlier to show a more granular view of the variance for the years open for alcohol offerings, price ranges, different cultures and specialties respectively

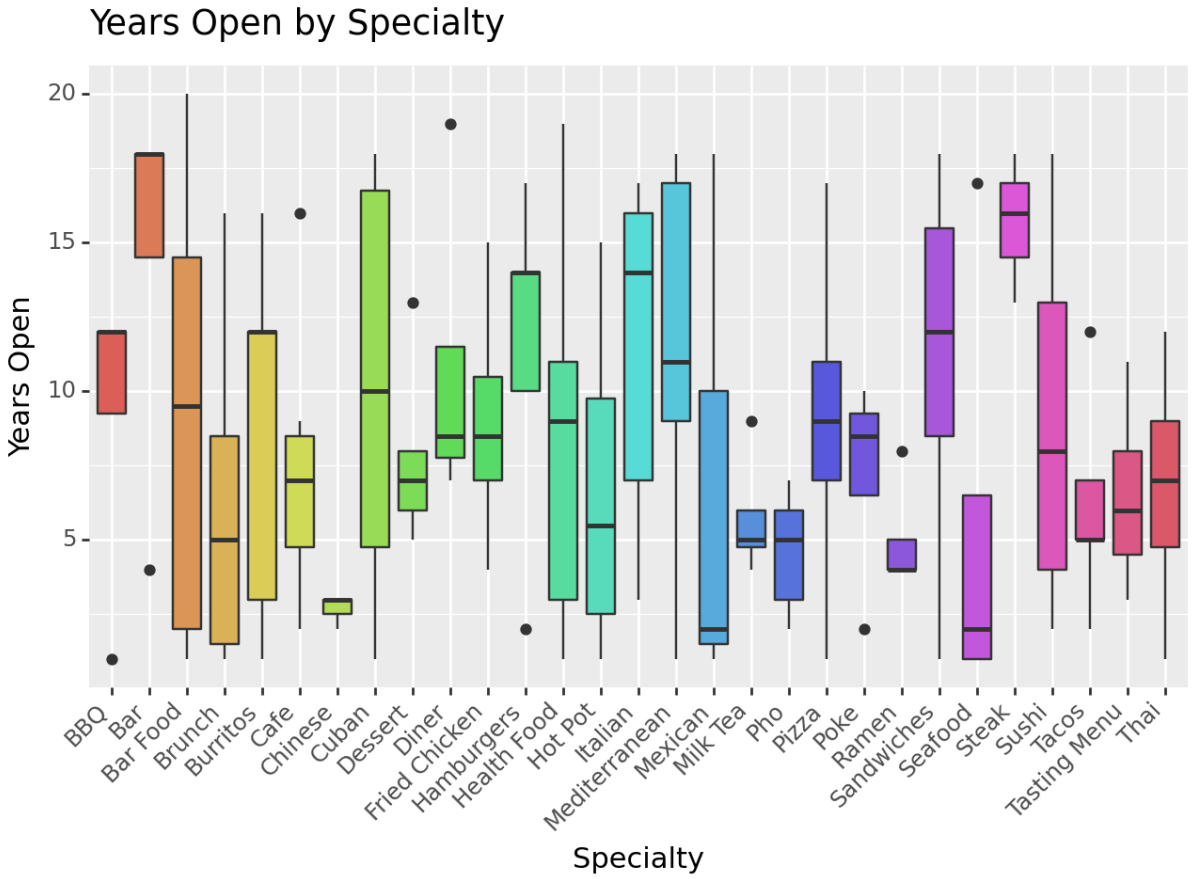



Years Open by Alcohol Availability
$1-15
$100+
$15-50
$50-100
Years Open
20
15
10
5
Beer
Full
None
Wine
Alcohol Availability
AlcoholAvailability
Beer
Full
None
Wine


Years Open by Culture
Years Open
20
15
10
5
American
Chinese
Cuban
French
Hawaiian
Italian
Japanese
Mediterranean
Mexican
Mexican
New American
Thai
Vietnamese
Culture


Years Open by Specialty
Years Open
20
15
10
5
BBQ
Bar
Bar Food
Brunch
Burritos
Cafe
Chinese
Cuban
Dessert
Diner
Fried Chicken
Hamburgers
Health Food
Hot Pot
Italian
Mediterranean
Mexican
Milk Tea
Pho
Pizza
Poke
Ramen
Sandwiches
Seafood
Steak
Sushi
Tacos
Tasting Menu
Thai
Specialty

## Brief Discussion and Explanation:

Using a Gradient Boosting Tree and data preprocessing, this study examines the influence of culture, specialty, price, and alcohol availability on restaurant longevity. The analysis provides valuable insights into how these factors collectively shape the operational lifespan of restaurants. Basically we looked at things like the type of food they serve (like Italian or Chinese), what special dishes they're known for, whether they serve alcohol, and their price ranges. We make sure our information is complete and use a method called "OneHotEncoder" to organize it neatly. Then, we use a special technique, known as a Gradient Boosting Tree, to analyze how these factors affect how long a restaurant stays open. We split our data into two parts: one for building our model (80%) and the other for testing it (20%). Finally, we create several charts to visually show how long restaurants stay open based on these factors, excluding any unusual cases to get a clearer picture.

Based on these results, it seems that in general restaurants with some type of alcohol offering seem to be in operation the longest. In addition, American, Cuban, Italian, and Mediterranean culture restaurants seem to all have the longest years in operation, with each of these cultures having an average lifespan of over 10 years. Finally, it seems that in terms of specialty, bars and steak focused restaurants and businesses have the longest operational lifespan with both having an average of over 15 years, though there is the obvious requirement of being licensed by the state in order to serve alcohol.

## Q2 (Clustering) Can we identify distinct clusters of restaurants near Chapman University based on their combination of predictors: Rating, HealthInspectionRating, DistancefromChapman and AverageMealPrice?

### Variables Involved:

- Rating (Continuous)
- DistancefromChapman (Continuous)
- AverageMealPrice (Categorical)

### Cleaning:

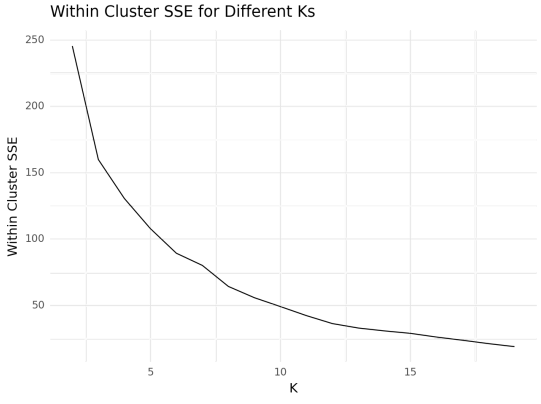
Missing values will be dropped, and reindexed. Standard Scale continuous variables and make Avg Meal Price into numerical categories

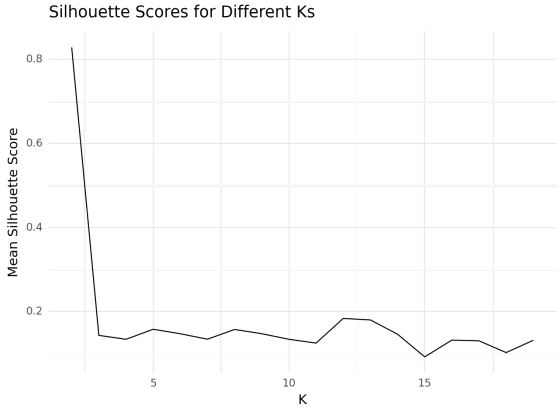
### Modeling/Computation:

Make two clustering models K-Means, and Gaussian Mixtures

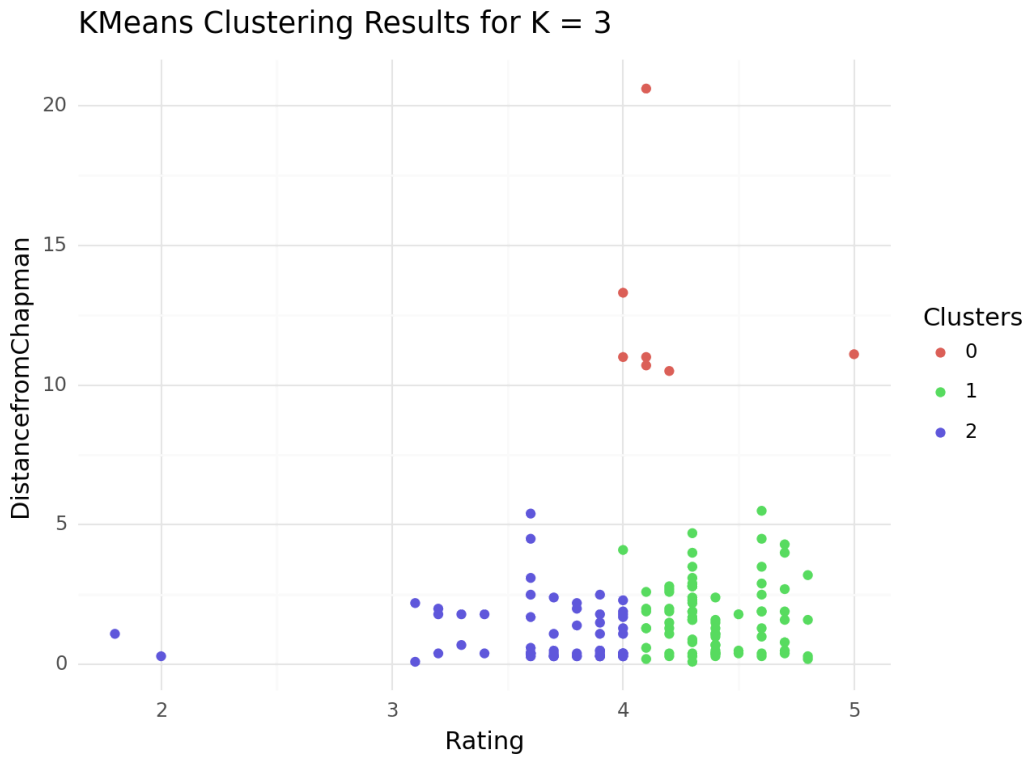
- Use the elbow method for K-Means to find the optimal number of clusters.
- For Gaussian Mixtures I use Bayesian Information Criterion (BIC) to choose the optimal number of components (clusters) based on the trade-off between model fit and complexity

First, we tidy up our data by removing any entries that have missing values and adjusting the numbers to make them comparable. Then, we use two different clustering methods, K-Means and Gaussian Mixtures, to find these groups. For K-Means, we use a technique called the 'elbow method' to determine the best number of groups, and for Gaussian Mixtures, we use something called the Bayesian Information Criterion to balance the model's detail and simplicity. We also create graphs showing different potential groupings to help us decide the best number of clusters for KMeans.

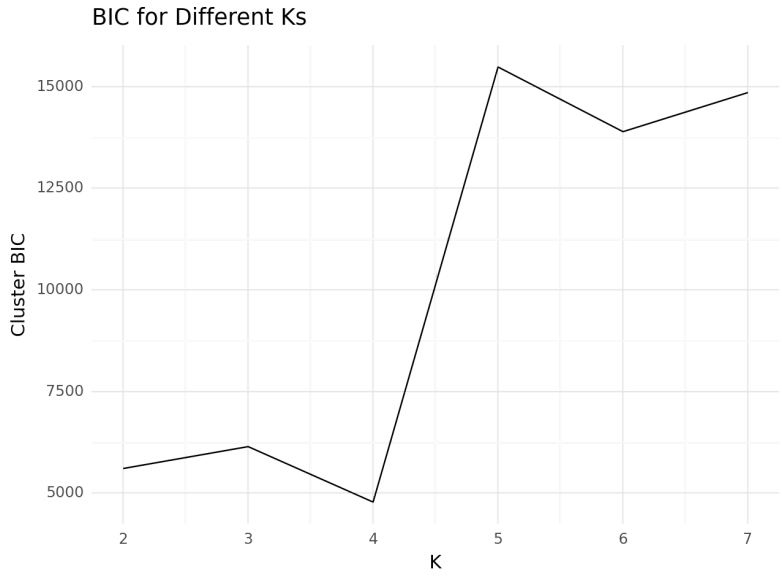


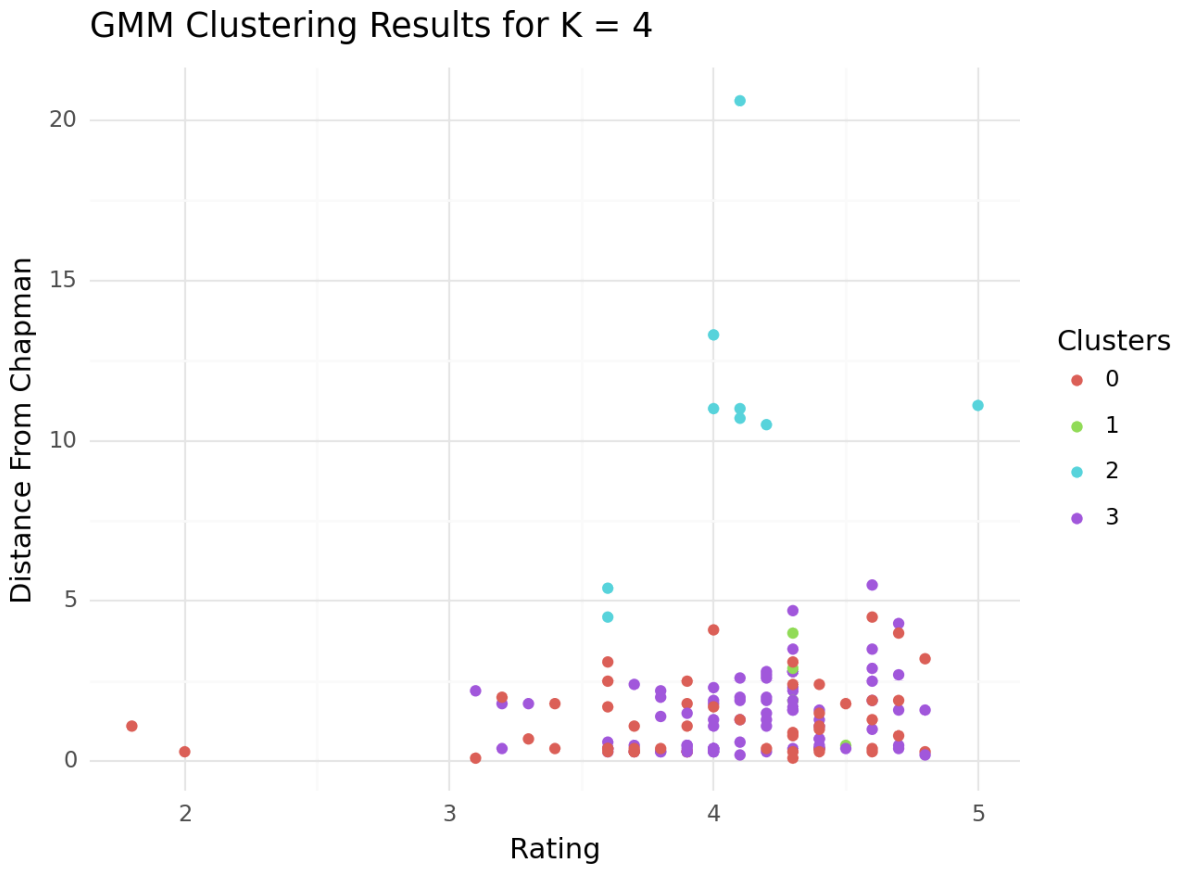

Looking at these graphs, the ideal cluster size for K-Means would be 3 clusters




Using K-Means, these are the 3 clusters we get, with Cluster 0 being the furthest from Chapman, but tending to have ratings that are generally above 4.0, Cluster 1 and 2 being roughly the same distance from Chapman, but with Cluster 1 having much better ratings between 4 and 5 while Cluster 2 has poorer ratings of between 3 and 4.




Looking at this BIC Graph, we will choose a cluster size of 4 for GMM. We use the elbow method to find the lowest BIC score, which corresponds to the best number of clusters. The elbow method is like finding the best number of groups for your data by looking for the point where adding more groups doesn't make your model much better.




We have 4 clusters using GMM, though GMM in general does not seem like an appropriate clustering method. As indicated within the graph, Clusters 0 (red) and 3 (purple) seem to be fairly interweaved, and its hard to grasp distinctive characteristics between the clusters.

**Brief Discussion:**

This analysis is effective at answering the question because clustering helps group data points that are similar to each other. It will help find natural grouping restaurants with similar ratings, health inspection results, distances from Chapman University, and average meal prices. This can help identify distinct categories of restaurants based on these features

## Q3: Impact of Culture, Specialty, and Reviews on Ratings - Analyze how a restaurant's cultural background, specialty food type, and number of reviews affect its ratings.

### Methods:

The variables we used were Culture (Categorical), Specialty (Categorical), and Reviews (Continuous) in order to predict Rating (Interval). For cleaning and preprocessing, we handle missing values, use OneHotEncoder() for categorical variables and use StandardScaler() to z score continuous variables. Basically, we clean our data by dropping any missing information and using a technique called "OneHotEncoder" to neatly organize different categories, and then use “StandardScaler()” to adjust numerical values so they're on a similar scale in order to compare these different types of information effectively. In terms of the model, we decided on a Random Forest supervised model, utilizing Grid Search to find ideal hyperparameters. Basically, we opted for a type of computerized analysis called Random Forest, which is really good at making predictions based on our restaurant data. To make sure it works its best, we used a technique called Grid Search, which is like trying out different settings to find the perfect combination for accurate results. To ensure accuracy, we'll divide our data, using 80% for training our model and 20% for testing it. The model will then predict the rating of restaurants based on these factors.

### Results:

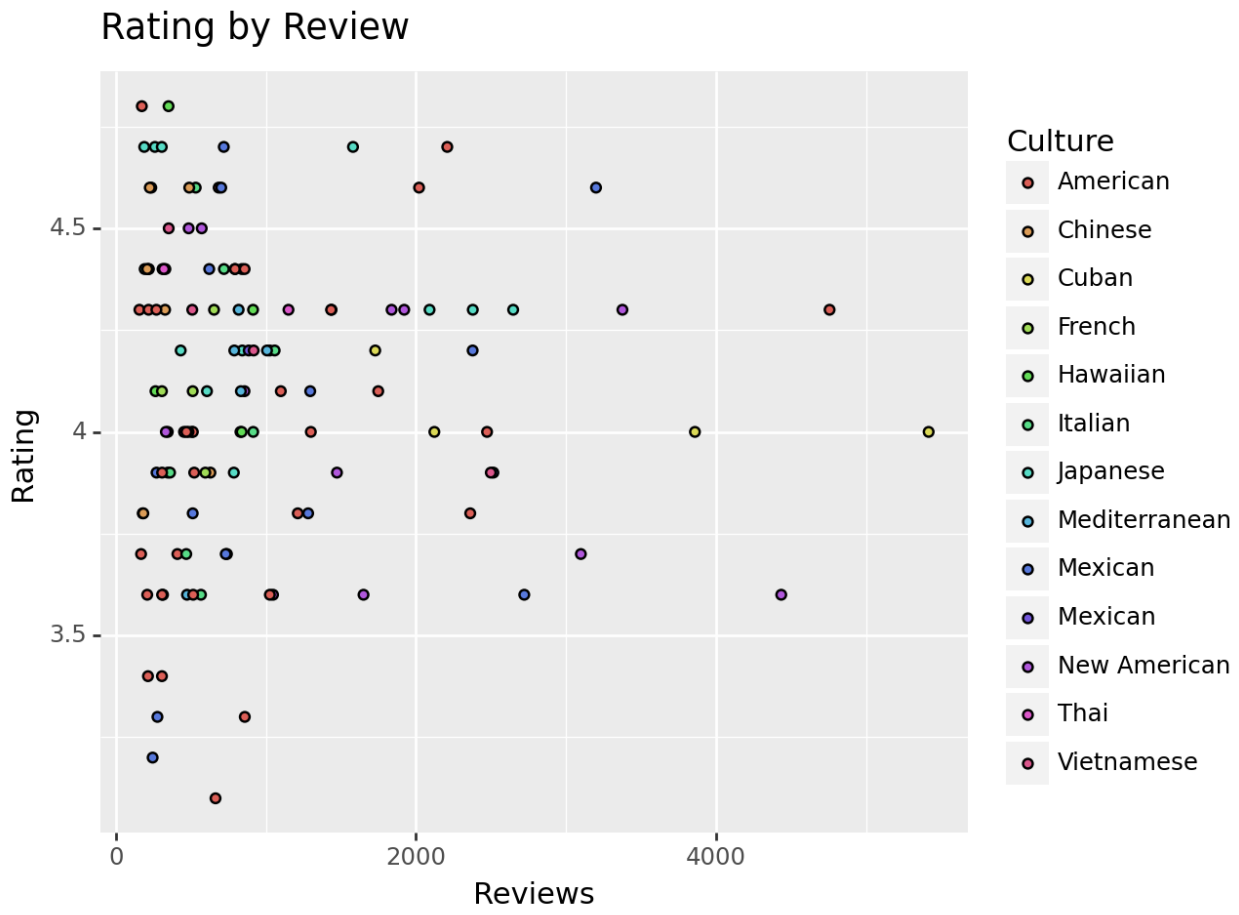
We experimented with several methods like Linear Regression (both simple and with added complexity called polynomial features), Ridge and Lasso (which are tweaks to Linear Regression), and a few others like K-Nearest Neighbors, Decision Tree, and Gradient Boosting Tree. However, these methods didn't work out too well because they were either too complex or too simple, causing issues like overfitting, where the model is too tailored to our specific data and doesn't apply well to new situations. In the end, Random Forest was our best choice as it balanced complexity and accuracy better than the others, making it the "least bad" option for our needs.

| | Training Set | Testing Set |
|---|---|---|
| MSE | 0.0409 | 0.1868 |
| MAE | 0.1587 | 0.3415 |
| MAPE | 0.0407 | 0.0896 |
| R2 | 0.8176 | 0.1425 |

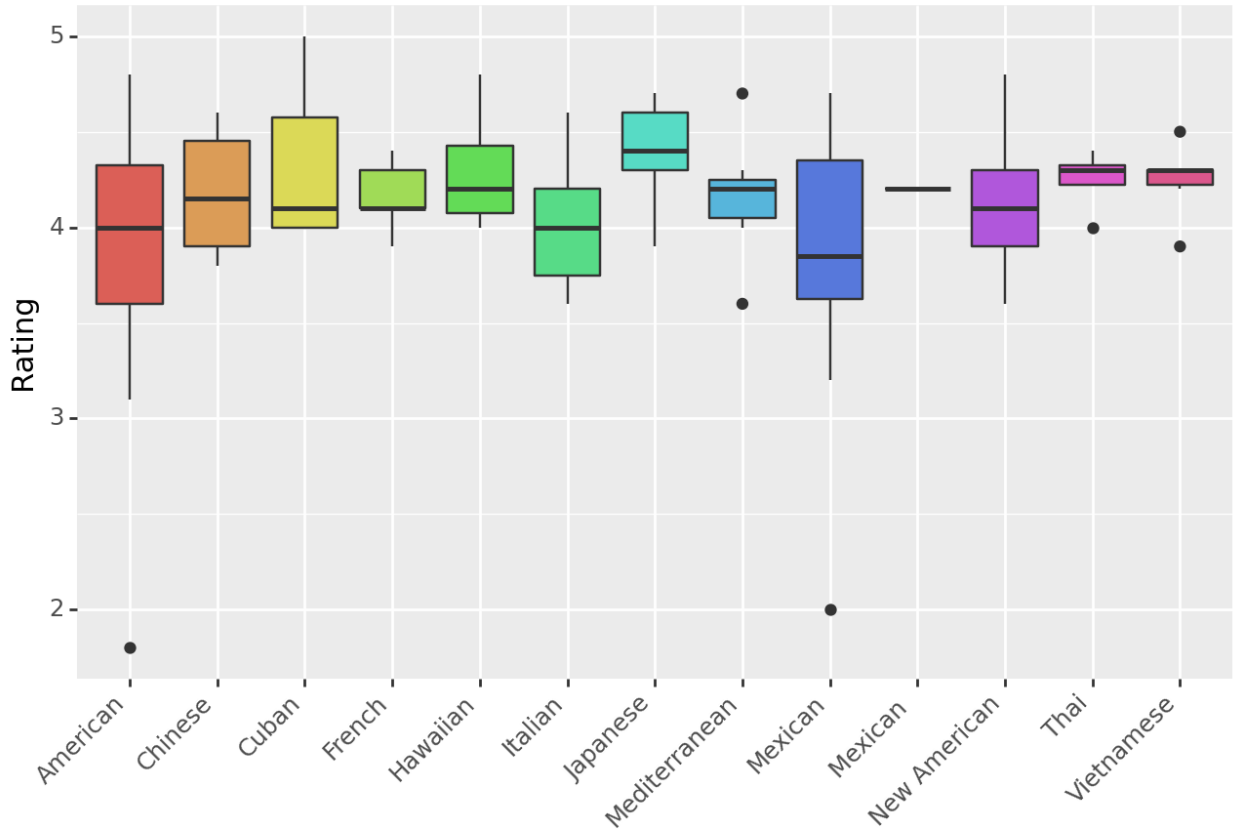
In our Random Forest model, we use four key measures to check its accuracy. MSE (Mean Squared Error) and MAE (Mean Absolute Error) tell us how far off the model's predictions are, with lower numbers being better. Our model does great on the training data but is less accurate on new, unseen data. The R2 score, which is like a grading system, shows our model fits well with the training data but not as closely with the test data, suggesting it's better at handling familiar situations than new ones.

**Graphs:**

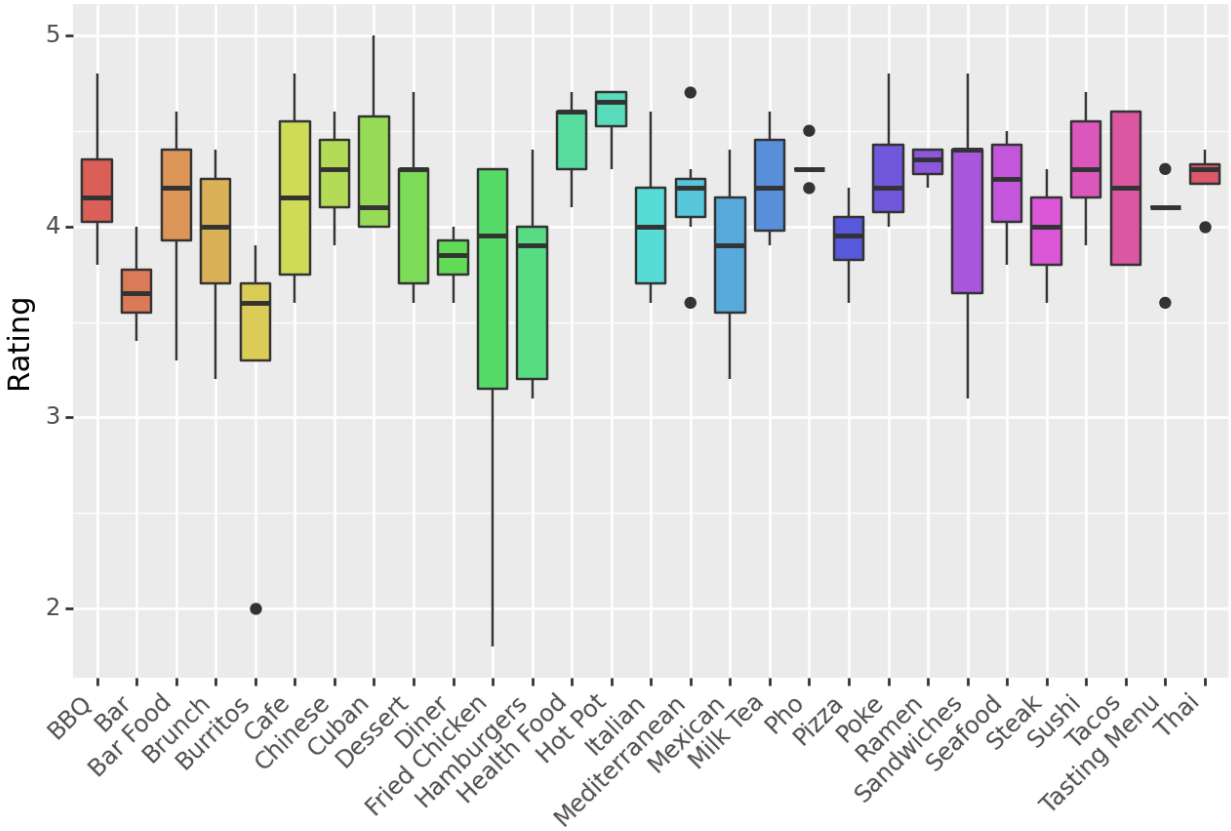
- Scatterplot that shows Reviews on X Axis, and Rating on Y Axis
- Box plot that shows the Rating for each Culture
- Box plot that shows the Rating for each Specialty



## Rating by Culture




## Rating by Specialty



Looking at the first scatterplot, it's interesting how when a restaurant gets more reviews, it seems to converge at around the low 4.0 area. This could be explained by the fact that a restaurant with more ratings does tend to be more popular, which may attract more people that are potentially critical of the restaurant. The 2nd and 3rd boxplot graphs show that the typical rating for all cultures and specialties tend to hover around the 4.0 range, with cultures having a variance that is slightly above 4.0 and specialties having a higher variance around 4.0. There do seem to be some outliers at around 2.0, but in general given that the dataset is pulled from Yelp ratings I'm not overall surprised that they tend to be fairly high.

### Brief Discussion and Explanation:

We utilize a supervised random forest model to investigate how a restaurant's cultural background, specialty food type, and number of reviews influence its rating. The analysis, visualized via a scatterplot and 2 boxplots, reveals how different cultures and specialties correlate with customer ratings, providing valuable insights into customer preferences and satisfaction trends.

## Q4: (Clustering) How does the proximity to Chapman University affect restaurant success?

### Variables:

- Distance (Continuous)
- Success (Continuous) = Rating * Reviews via Feature Engineering

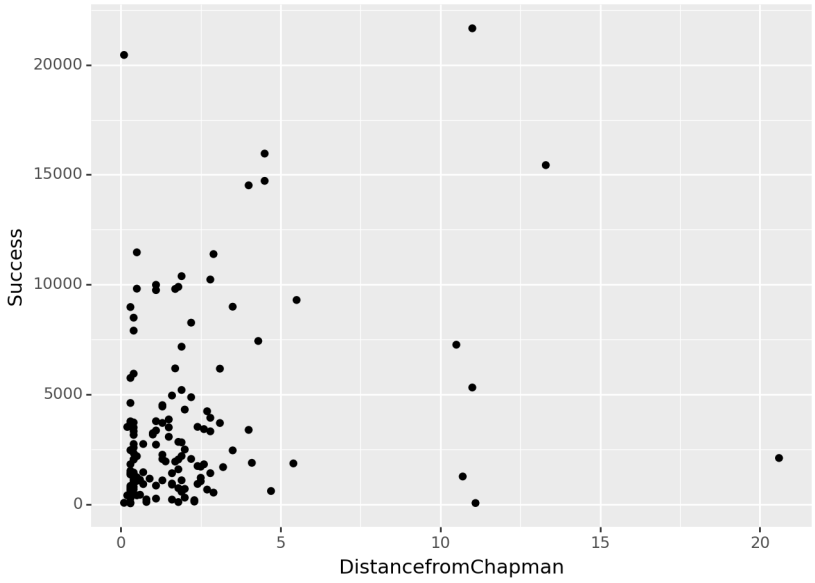
### Cleaning and Preprocessing:

- Handle missing values.
- StandardScaler() to Z Score continuous variables

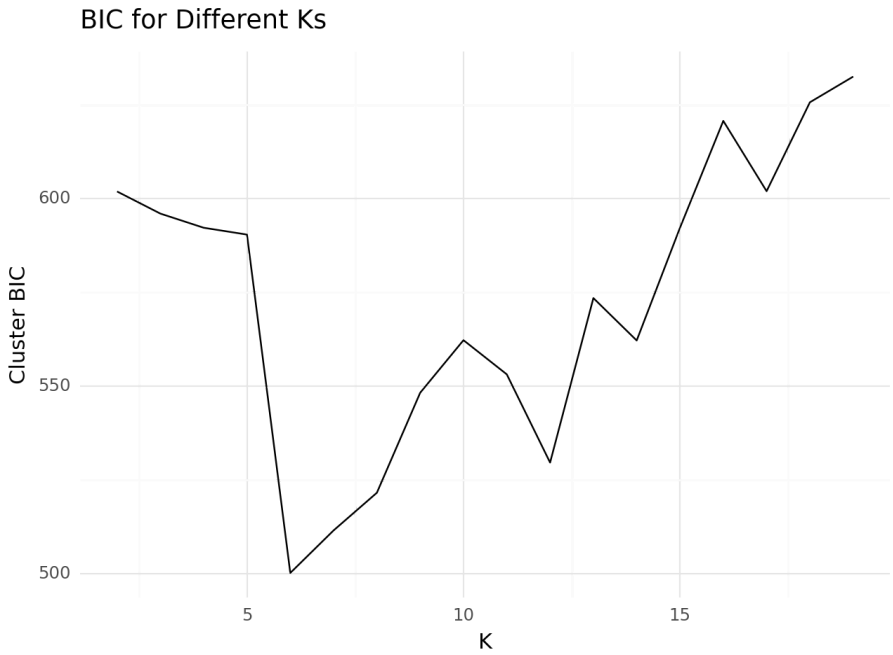
### Modeling/Computation:

- Apply a clustering algorithm like K-Means or Hierarchical Clustering. (Or other depending on how the data is and which would be appropriate)
- Define 'success' as high ratings and focus the analysis on restaurants meeting these criteria.

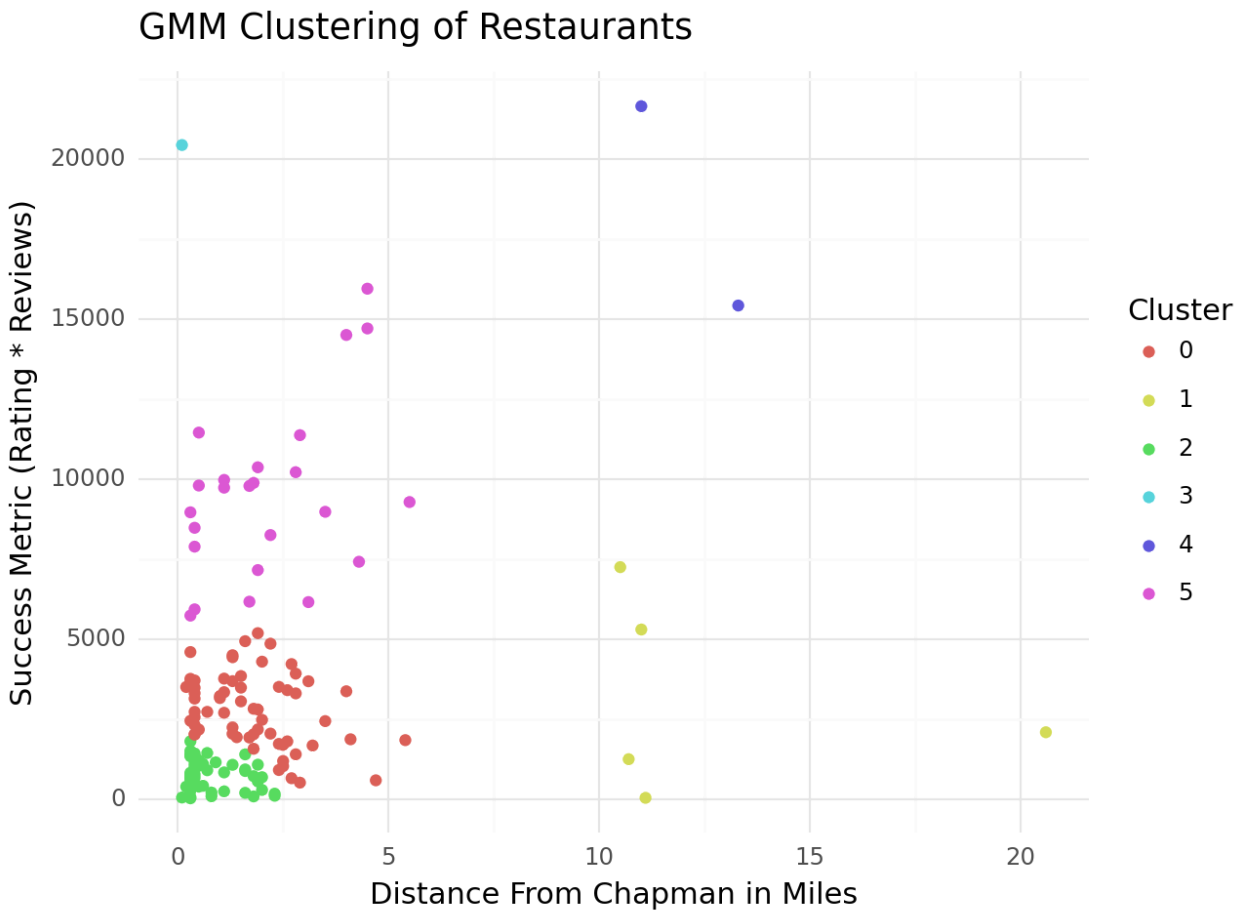
We're using a clustering method to group restaurants based on how close they are to Chapman University and how successful they are, measured by customer ratings and the number of reviews they get. We first clean up our data, ensuring there are no missing pieces and adjusting the numbers to a common scale. Then, we use a clustering technique like K-Means or Hierarchical Clustering in order to see if there are any patterns related to the distance from Chapman University and the success metric.



Looking at this graph, it does seem like DBSCAN or GMM would be the two most appropriate clustering methods. I would specifically argue for GMM, as while there is noise, or outlier points, there does seem to be varying density through the clusters, or how many points are in a set location, and the cluster shapes do seem to be roughly spherical in nature. There also seems to be a fair amount of overlap between clusters, which is another strength of GMM. GMM, short for Gaussian Mixture Model, is like a clustering method that can spot subtle patterns and group similar things together in data, and it's especially good because it can handle complex, overlapping groups more effectively than simpler methods. Because we are using GMM, we choose the number of clusters using the elbow method to find the BIC from the graph below.




The best model, according to BIC, is the one with the lowest score. It's like finding the best set of instructions for your puzzle: detailed enough to be accurate, but not so complex that it's hard to follow. We use the elbow method to find the lowest BIC score, which corresponds to the best number of clusters. The elbow method is like finding the best number of groups for your data by looking for the point where adding more groups doesn't make your model much better. It's about finding the sweet spot where the model is accurate but not too complicated. The cluster size we choose will be 6.




## Brief Discussion and Explanation:

To surmise the findings of the graphs, first let's talk about the Silhouette Score of 0.3771, which is arguably pretty good. The Silhouette Score is a way to measure how well our computer model has grouped similar things together in the data. In addition, it is imperative that the restaurants are close to Chapman. Being within walking distance is not important but being within a certain valence will be beneficial to the restaurant in the long run. In looking through the data once clustered, the optimal segmentation was 3 groups. Cluster one which is labeled 0 (in red) is mainly Old Towne Orange and some offshoot restaurants that did not necessarily fit the description of another cluster. Cluster 1 (green) is restaurants that are mainly past Glassell and Katella. The final cluster (blue) are restaurants that are within range but have such a high success metric and are so well established that it would be foolish to segment them with cluster 0.

## Q5: (Supervised Model and Dimensionality Reduction via LASSO) Influence of Meal Price and Alcohol Availability on Ratings

### Variables:

- Average Meal Price (Categorical)
- AlcoholAvailability (Categorical)
- Ratings (Interval)

### Cleaning and Preprocessing:

- Handle missing values.
- OneHotEncoder() for Categorical.

### Modeling/Computation:

- Use a supervised model with dimensionality reduction to predict complex relationships such as a Ridge or Lasso
- Train/Test split of 80-20 for model validation.

We are using a model to understand how the price of meals and whether a restaurant serves alcohol affect its ratings. To do this, we first clean our data, dropping entries with any gaps and organizing categories like meal prices and alcohol availability using OneHotEncoder(). Then, we use a special approach called LASSO, which helps simplify the data we have, making it easier for our model to understand complex relationships. We chose the 'alpha' for Lasso, which is a key setting that determines how much the model focuses on the most important factors, by testing different values to find the one that gives us the best balance of simplicity and accuracy in our analysis. We tested different models, including Linear Regression with extra features, Ridge, KNN, and a few tree-based methods. However, LASSO turned out to be the most effective, even though it wasn't perfect, it gave us the clearest insights compared to the others. We split our data, using 80% for building the model and 20% for testing its accuracy.

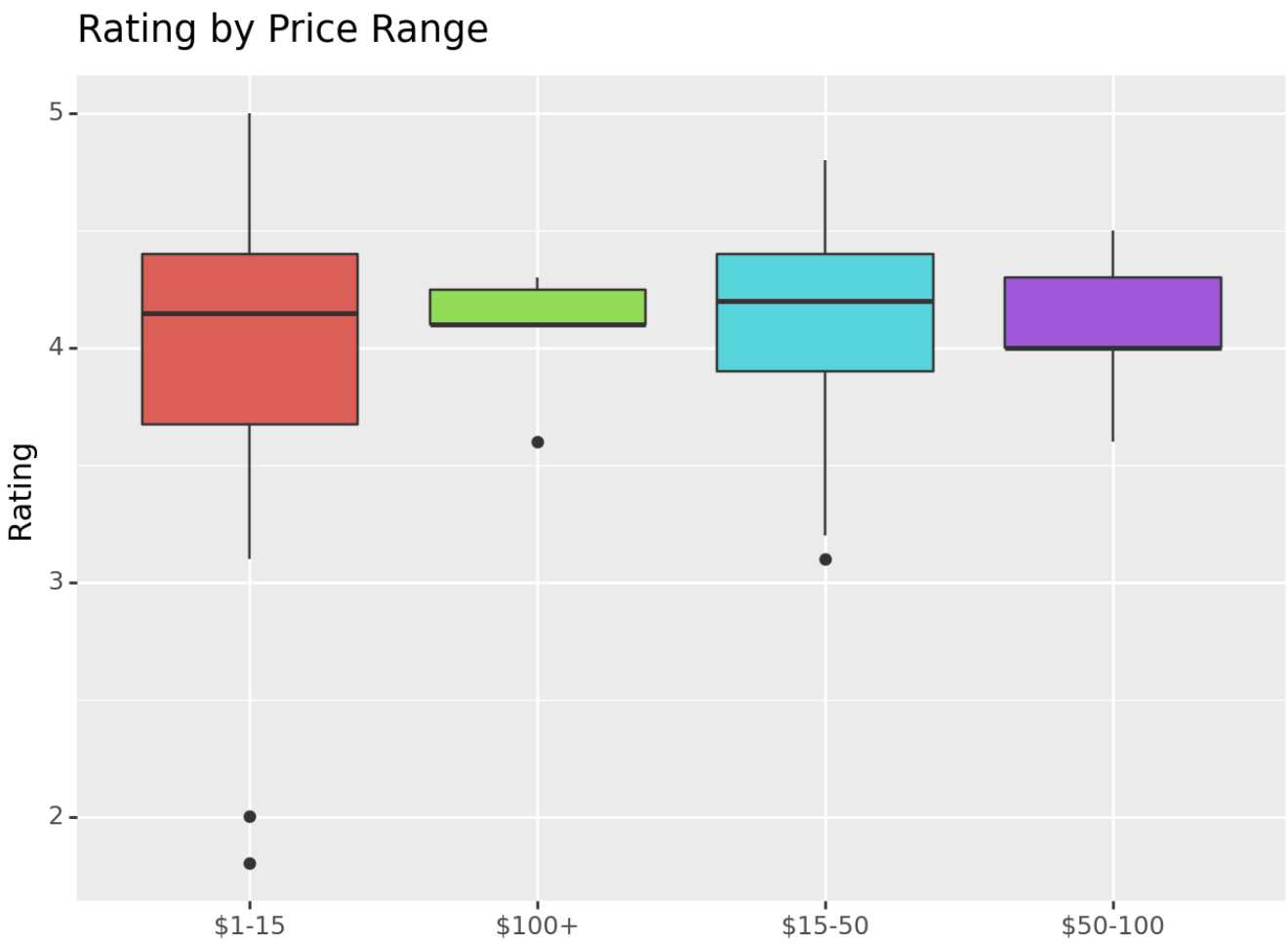
**Results**:

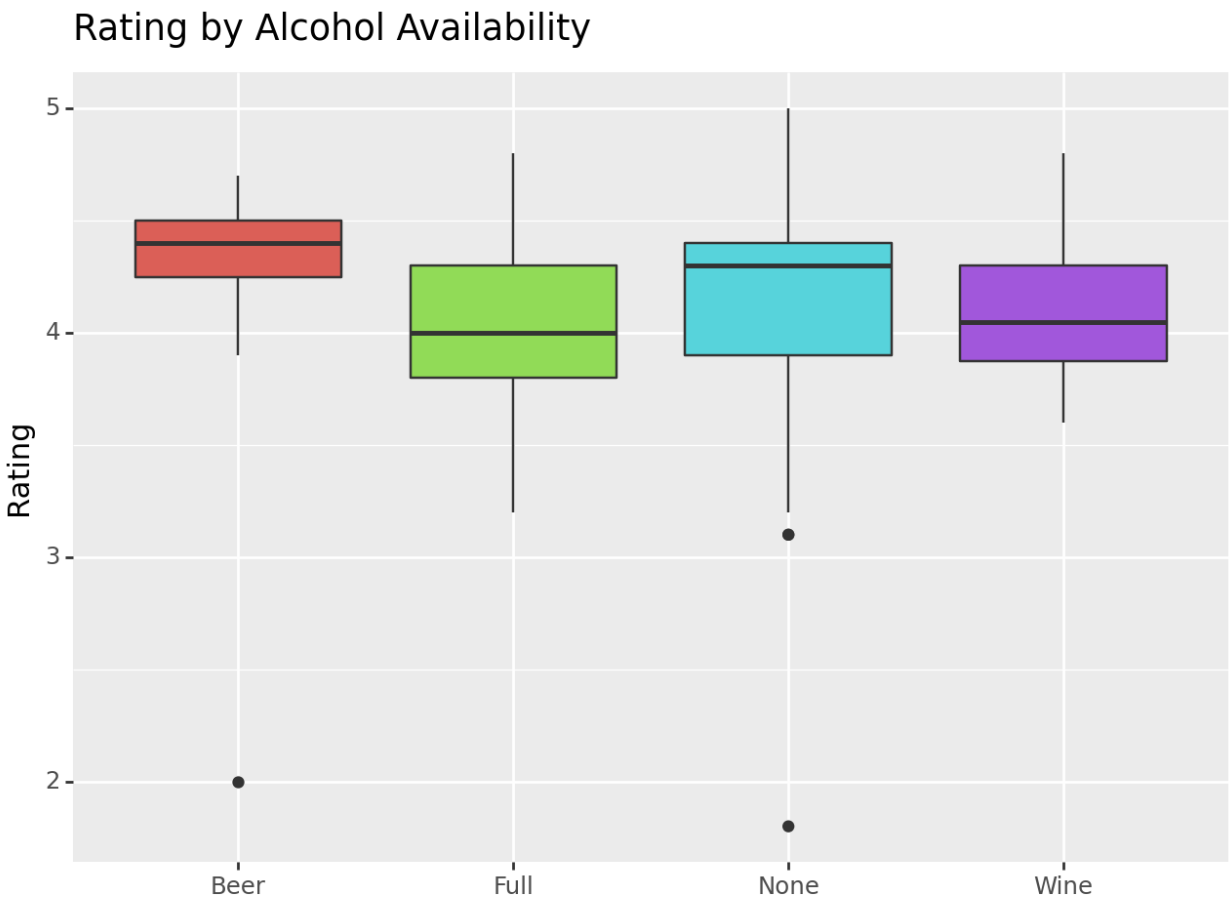
| | Training Set | Testing Set |
|---|---|---|
| MSE | 0.2263 | 0.1544 |
| MAE | 0.3436 | 0.3207 |
| MAPE | 0.0952 | 0.08 |
| R2 | 0.0278 | 0.1008 |

The outcomes of our model weren't outstanding, but it's important to point out that when we used Linear Regression with a method called Lasso and K Fold Validation, it actually did better than the standard Linear Regression, Ridge, K Nearest Neighbors, and other tree-based models at our disposal. This means that even though it wasn't perfect, the Lasso approach was the most effective at making sense of our data compared to the other methods we tested. Lasso is like a smart filter for data analysis; it helps focus on what's truly important by reducing the less significant details. K Fold Validation, on the other hand, is like a thorough test for the model, dividing the data into several parts, using some for learning and some for testing, to ensure the model is reliable and works well in different scenarios.

**Graphs:**

- Box plot that shows price ranges on x axis and rating on y axis
- Box plot that shows alcohol availability on x axis and rating on y axis






## Brief Discussion and Explanation:

Looking at the first graph, it seems that except for a few outliers, each of the 4 price ranges tend to have an average rating of around 4.0. That being said, the $ or $1-15 price range does seem to have the highest variance in its rating. In terms of the 2nd graph, restaurants that serve beer only tend to have the highest ratings, while interestingly enough restaurants with no alcohol offerings have the second highest average.

This study employs a Linear Regression model with Lasso Dimensionality Reduction to explore the complex relationship between average meal price, alcohol availability, and customer ratings. This analysis can provide insights into how pricing strategy and alcohol influence customer satisfaction, aiding in strategic decision-making.

## Q6: Feature Importance on Average Meal Price

## Examine most important features that predict average meal price at a restaurant near Chapman University

### Methods:

We are using Restaurant Type (Categorical), Culture (Categorical), Specialty (Categorical) DistancefromChapman(Continuous), Rating(Continuous), Reviews(Continuous), CompetitorDensity(Continuous), Alcohol Availability(Categorical), and YearsSinceOpen(Continuous) being the features, and AverageMealPrice as the target variable denoting the average meal price category. The AverageMealPrice column is then mapped to numerical categories using the defined 'price_mapping' dictionary, assigning values 1 through 4 corresponding to the '$ ' to '$$$$' price categories

Categorical variables in the features are preprocessed using OneHotEncoder, resulting in a new DataFrame 'X_encoded' that includes binary columns for each category. Subsequently, the dataset is split into training and testing sets with an 80-20 ratio using the train_test_split function from scikit-learn.

Finally, a Random Forest Regressor model is instantiated and trained on the training data. The model is configured with 110 estimators, aiming to predict the average meal price category based on the specified features. This approach allows for the exploration of relationships between restaurant characteristics, including type, culture, specialty, alcohol availability, and various numerical factors, with the ultimate goal of predicting meal prices through the Random Forest model.

### Results:

We chose to employ a Random Forest Regressor due to its superior performance in feature selection. In our exploration of predicting average meal prices at restaurants near Chapman University, the selected Random Forest Regressor demonstrated commendable performance on the training set, evidenced by a low Mean Squared Error (MSE) of 0.0416 and a high R-squared ($R^2$) of 0.8892. These metrics suggest a robust

fit to the training data, indicating the model's ability to accurately predict average meal prices based on the specified features.

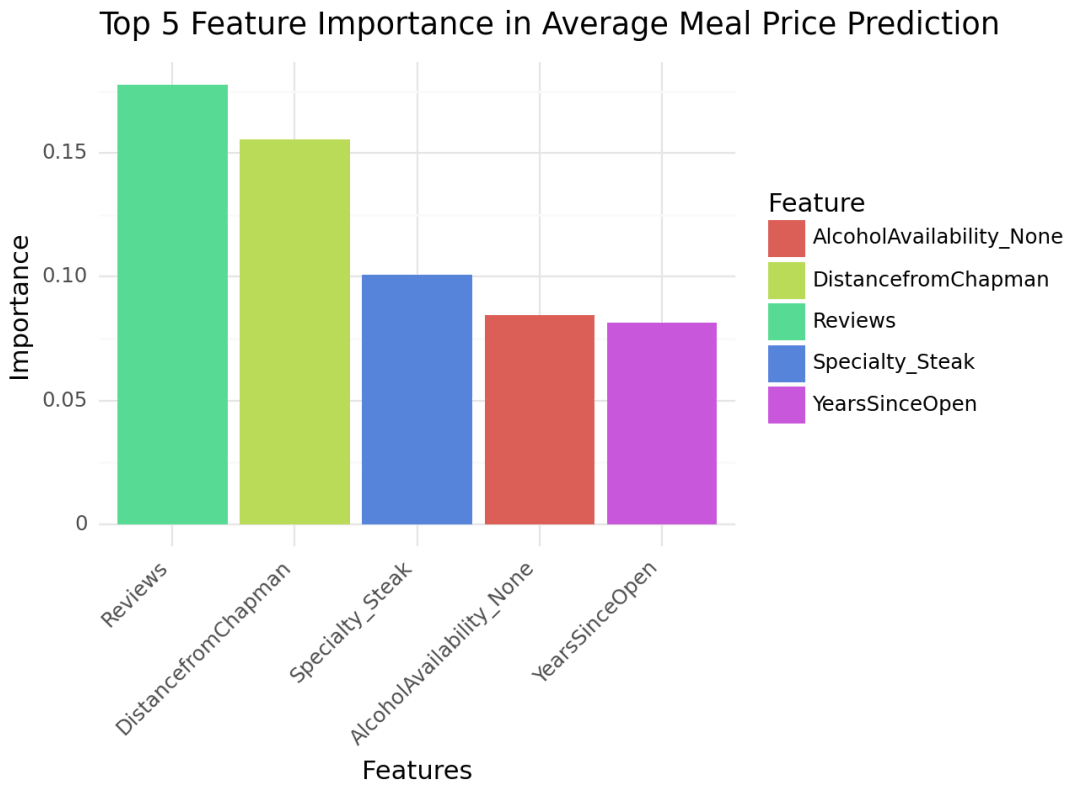
However, upon evaluation of the testing set, a noticeable decrease in performance was observed, with a higher MSE of 0.1500 and a lower R-squared of 0.3389. This discrepancy between training and testing results implies potential overfitting, emphasizing the need for further refinement to enhance the model's generalization to new, unseen data.

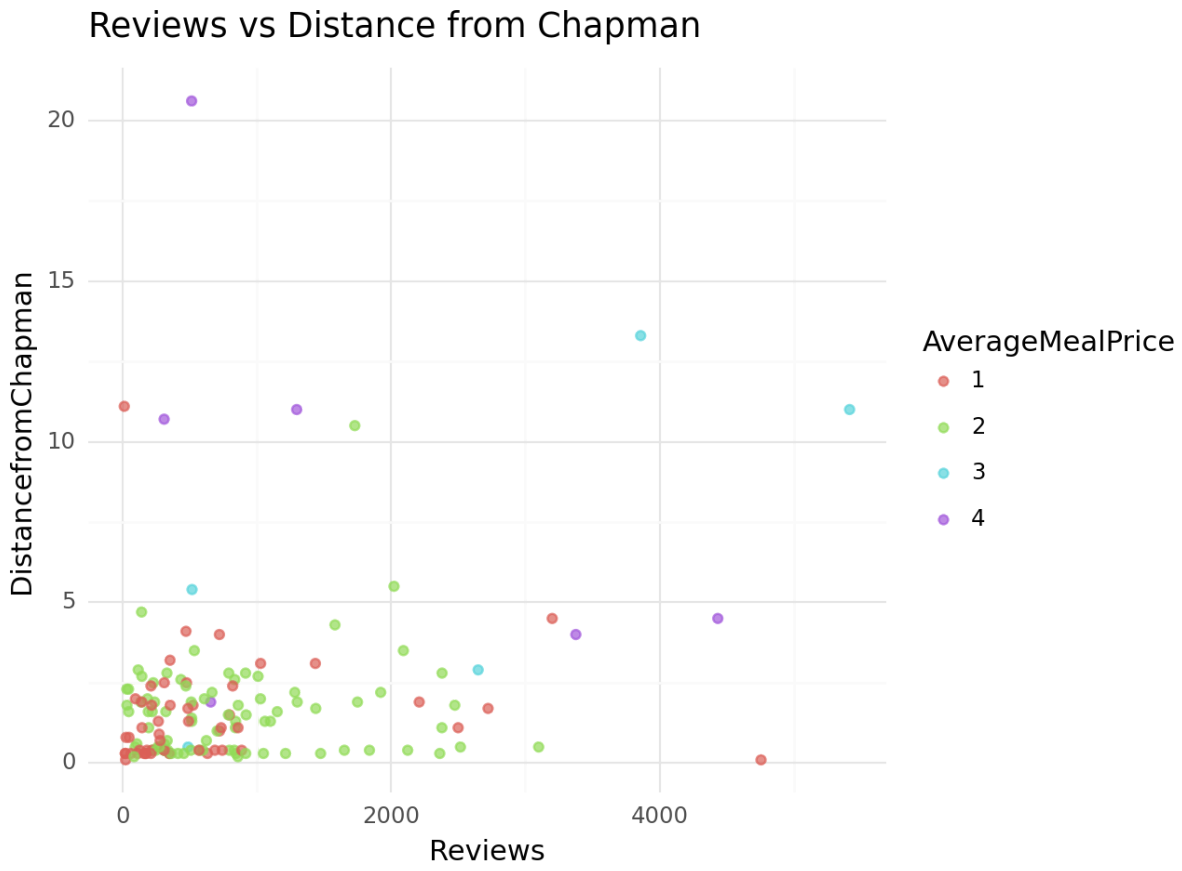
Despite this, the feature importance analysis identified key predictors, including 'Specialty_Tasting Menu,' 'AlcoholAvailability_None,' 'Reviews,' 'DistancefromChapman,' and 'Rating.' These features play crucial roles in influencing average meal prices. While the Random Forest Regressor has shown promise, addressing overfitting and fine-tuning the model parameters will be essential for optimizing its performance and ensuring reliable predictions in real-world scenarios.

Among these, customer reviews emerged as the most impactful, contributing 17.78% to the model's decision-making process. Proximity to Chapman University, indicated by 'DistancefromChapman,' followed closely with a significance of 15.55%. The specific restaurant specialty, particularly in the case of steak ('Specialty_Steak'), played a noteworthy role, contributing 10.08%. Additionally, the absence of alcohol availability ('AlcoholAvailability_None') and the number of years a restaurant has been in operation ('YearsSinceOpen') were identified as influential factors, each contributing 8.45% and 8.16%, respectively. These findings emphasize the crucial role that customer sentiment, location, specialty offerings, alcohol availability, and the longevity of a restaurant play in determining average meal prices. Understanding these key factors can guide decision-making and optimization strategies for restaurant operators in the vicinity of Chapman University.

**Graphs:**

- The first plot is a Bar plot showcasing the top 5 features with the highest importance in predicting AverageMealPrice
- The second plot is a detailed scatter plot depicting the relationship between the most important features Reviews and the DistancefromChapman. The points in the scatter plot are colored according to the AverageMealPrice, providing a visual representation of how these variables interact






**Brief Discussion:**

In our study to figure out what affects meal prices at restaurants near Chapman University, we looked at a mix of different factors like the type of restaurant and customer reviews. We made sure our data was in good shape by dealing with missing info and getting it ready for our analysis. Then, we used a smart tool called a Random Forest model to help us find the most important features in predicting meal prices.

To make sense of our findings, we used cool charts and graphs made with ggplot. One graph showed how the distance from Chapman University vs Reviews connects with meal prices. Another chart pointed out the top 5 things that really matter in predicting meal prices.

These visuals aren't just for looks—they help us understand the relationships between different factors and how good our predictions are. With these easy-to-read charts, restaurant owners can make smarter decisions. For example, they can see how the distance from Chapman University affects prices and use that info for setting prices or planning marketing strategies. The competitiveness metric we came up with can also help them understand how they stack up against others in the area. All in all, these

simple yet powerful visuals from our analysis help businesses make informed choices and stand out in the competitive restaurant scene around Chapman University.

Based on these results the biggest factors that affected were the amount of reviews, distance from Chapman, alcohol availability, and years since opening. With these known features this can help restaurants better price their food and help them be able to charge more for food by providing these steps. This can also help prospective restaurant owners looking in the Chapman area to make a better decision on what important factors affect the prices of meals in their restaurants.

Ultimately, the predictive model's insights, coupled with strategic marketing actions informed by these findings, can help local restaurants near Chapman University better connect with their target audience, optimize pricing strategies, and ultimately enhance their competitive position in the market