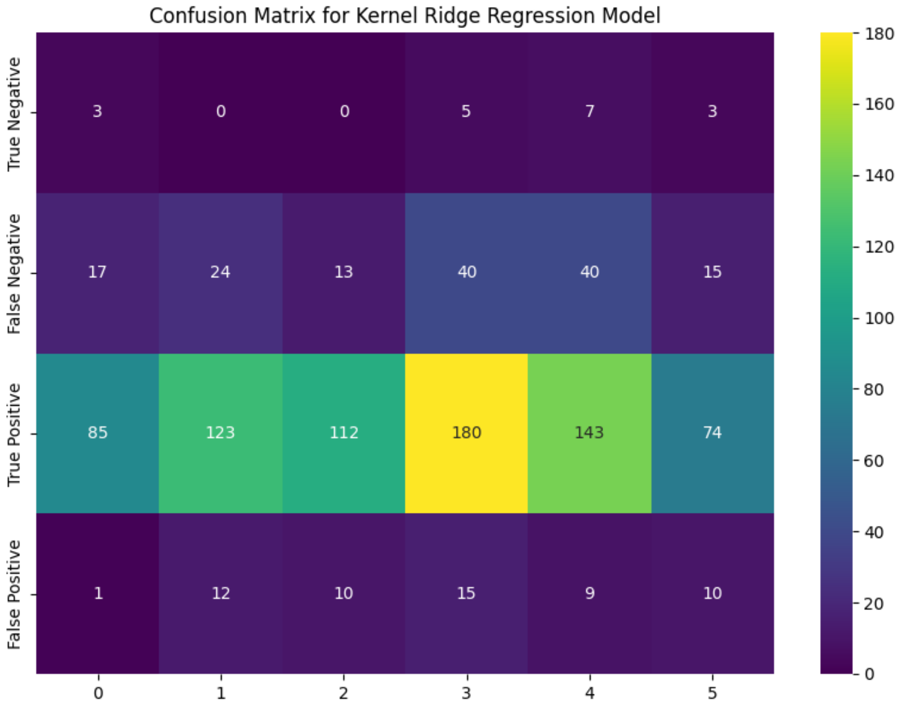
Data Visualization Report:

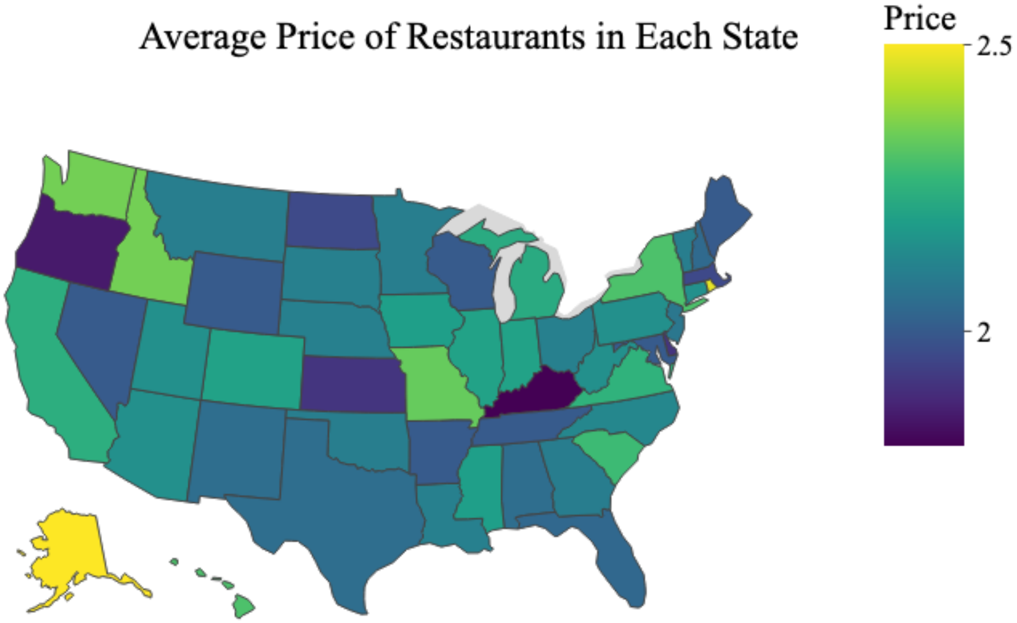


We chose to create a heatmap as our project revolved around a classification task to predict a restaurant's rating between 0 and 1. The rating was rounded to a whole number to fit this heatmap. This visualization allows us to see how well our model is performing and how accurately it predicts the correct and incorrect labels.

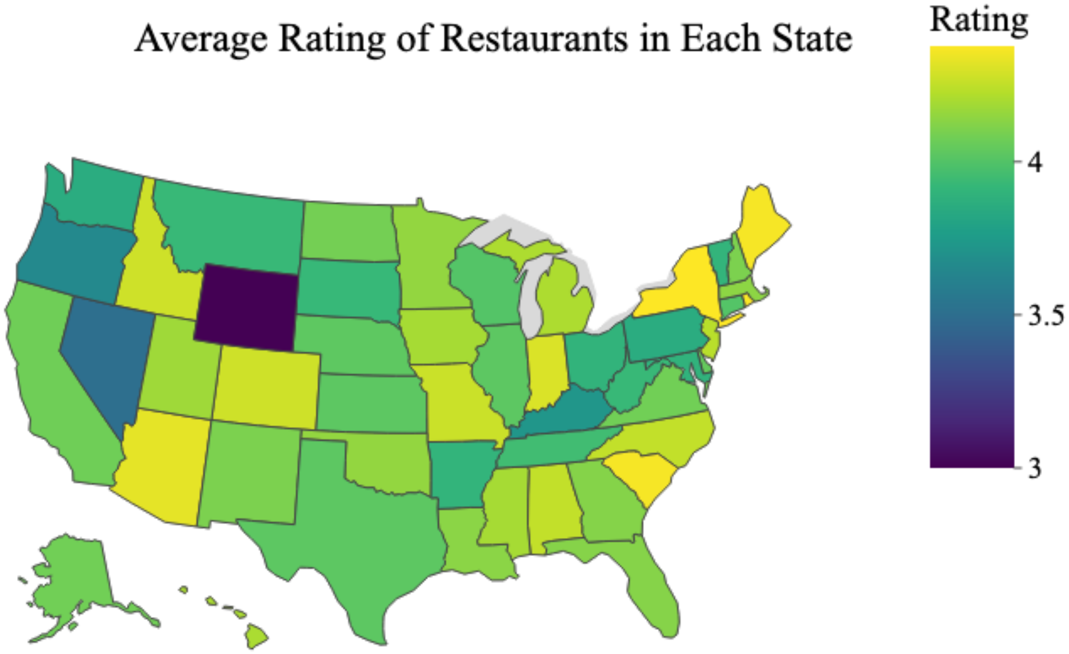
Alternatively, we could have used a histogram or a line chart to see how accuracy improved as we trained the model. We could have used a pie chart to portray the same data. Although these alternatives exist, we believed that a heatmap would best allow us to compare the true negatives and true positives, while considering the different labels. As we can see most of our data had a rating of 3, and this is what which is the average of the possible ratings.

Some challenges of creating this visualization was the fact that this was not a binary classification task, but rather a classification task with 6 labels. Figuring out the best way to display the information among the 6 classes and what each row represents was a challenge in creating this visualization.

This is recognizably a heatmap, however, context would be needed in order to better understand what the visualization is showing. Although we have x and y labels, we do not know that they are the ratings and what was used to predict these ratings.



Average Price of Restaurants in Each State



Average Rating of Restaurants in Each State

To better understand our data and our hypothesis, we choose to display our data in the form of choropleths. We created two choropleths, one with the average ratings and one for the average price points for each state belonging to the United States of America. This visualization allows us to see how the region, location, and distance from the ocean affects the ratings and price point of restaurants in each state.

Alternatively, we could have communicated this result through a bar chart or pie chart. However, an advantage of choropleths is the fact that they relay positional information, something we are testing and is integral to our project.

There were not too many challenges in creating this visualization. The only issue that arose was that we needed to take the average ratings and prices of restaurants within a state. However, we expected to see restaurants closer to the ocean to have a higher rating and price point, but this is more difficult to convey within this choropleth without a larger investment of time.

This visualization is standalone and no context is needed due to the title and legend

Socio-historical and Ethical Report:

Socio-historical:

The historical context as well as current policies regarding seafood production, distribution, and consumption could impact the availability and pricing of seafood in certain areas, impacting our data analysis. Government regulations on fishing quotas, import/export tariffs, and subsidies could influence the supply and cost of seafood. For example, Alaska has a large seafood industry, and there are policies to promote sustainable fisheries, such as limiting the catch of certain species, restricting fishing gear, and running a seafood certification program ("Everything You Need to Know"). Seafood in states with stricter regulations might be more expensive, since it puts a limit to the amount of seafood available and might be more costly to catch fish. However, these regulations may result in higher quality seafood (e.g. certification program), which can result in a positive correlation between seafood quality, price, and rating. There are also social and cultural factors that should be considered. Coastal regions may have a stronger cultural connection to seafood cuisine, which could impact the overall popularity and rating of seafood restaurants, even if there were no correlation between price and distance from the ocean. We thought inland area seafood restaurants would be more expensive because of the transportation cost, which was not true. Lastly, economic inequality and urbanization may play a role in the pricing and rating of restaurants. States with more developed cities may have less affordable restaurants, while urbanized cities may lead to the clustering of higher-end restaurants with higher ratings. These more urbanized cities might have access to transportation to access seafood, which could be why we found the distance from the ocean to have a small impact on the rating of the restaurant. Although we did not look into these aspects, it would be crucial to do so, as there could actually be a correlation between urbanization of the location of the restaurants and the pricing of the restaurants, which raises a possibility to reframe our question. We should thus present our result with these external factors in mind.

Consumers may be interested in this correlation in considering what to eat, depending on the nature of the occasion (for example, if they are visiting a new area). Having such information available may convince users as to whether it's worth it to travel to another place to try a dish, or even whether it is worthwhile to try a restaurant in their area. The restaurants have a higher stake in this correlation, as such information may easily deter or promote their restaurant. Restaurants may want to consider their place and attempt to adjust their price points or the quality of their food accordingly, especially since the price level reflects the rating of the restaurant and the region where a restaurant is located leads to statistically significant differences in price. Additionally, seafood producers may have an incentive as their products are a large factor in this determination. They would likely hope that there is no correlation, as that would imply that their products and the ability of modern-day seafood transportation is advanced enough to present the ultimate quality of their items, regardless of location. Since there is a correlation between price and restaurant location, seafood producers might consider raising the prices of seafood for restaurants located in the regions with higher average seafood restaurant prices, which would benefit the seafood producers but harm the seafood restaurants.

Ethical:

We used the Yelp database; however, younger generations are drifting towards using other apps, such as Instagram, TikTok, and Beli, to search for restaurants in the area. Indeed, 35 to 54 year olds use Yelp the most, while 18 to 34 year olds use Yelp the least ("Yelp Demographics"). Also, 83% of Yelp users have reached at least college level education. The Yelp data used in this project may contain underlying historical or societal biases related to the demographics of the users (young college students or graduates) who leave reviews. We should also consider fake reviews, when restaurants often hire someone to write reviews or write favorable reviews themselves. Indeed, about 20% of Yelp reviews are fake, which would mean that the ratings of the restaurants are not an accurate representation of the restaurant quality or average customer satisfaction. This could result in a misleading interpretation of our data—restaurants with higher ratings are associated with higher quality and average customer satisfaction—which might not necessarily be true (Marie). Also, one of our assumptions when investigating whether or not price point affects rating was that the ratings for each restaurant are independent. Often, the same user rates different restaurants, so our assumption might not be true, which could make our data interpretation less accurate. Moreover, we analyzed the price point using $-$$$$$, instead of numeric price ranges. Since these price points were measured by Yelp, we cannot be certain if those price points are an accurate representation of the actual price of the restaurant, since the price and menu can change at restaurants. Lastly, for the ML models, we found that the accuracy of our results were not as high as we expected, which we believe was because of insufficient data for our models to accurately predict the rating of a restaurant from our features. We believe that if we had more data, we would have been able to achieve a higher r-squared score and a more accurate prediction, which could often lead to different results.

Our analysis could identify the location and behavior patterns of individuals who have left reviews for restaurants, revealing the socio-economic status of certain communities, potentially leading to discrimination or stigmatization. Especially with Yelp, the reviews are not anonymous, so the individual may be re-identified, even if personal identifiers are removed from the dataset. This loss of privacy could reveal the individual's place of residence, which could be associated with a certain socio-economic status as mentioned above. It is important to ensure that the data is properly anonymized and that any insights drawn from the analysis are only used to analyze the seafood restaurant industry in the area, rather than other community factors. A possible misinterpretation of our analysis is that it implies a causation between price and rating or between region and price. However, correlation does not necessarily mean causation, and there could be other factors that are affecting the results. As explained previously, some might assume the average price of the restaurant in an area reflects the socio-economic status of the community. With these misinterpretations, our data could be misused for discrimination and targeted advertising. Therefore, it is important to clarify that our results only show a correlation, and further research would be necessary to determine causation.

Works Cited:

"Everything You Need to Know about Alaska Fishing Rules and Regulations." *Alaskan Vacations: Fishing & Wildlife Tours*, www.pybus.com/blog/alaska-fishing-rules-and-regulations-bag-size-and-slot-limits. Accessed 11 May 2023.

Marie, Amanda. "Are Yelp Reviews Reliable? An Overview of Review Manipulation." *An Overview of Review Manipulation.*, 3 July 2022, blog.reputationx.com/are-online-reviews-reliable.

"Yelp Demographics: How Many People Use Yelp in 2023?" *The Small Business Blog*, 3 Apr. 2023, thesmallbusinessblog.net/how-many-people-use-yelp/.