Project Report

Transforming Yelp Reviews into Actionable Insights

Executive Summary

This project explores how Yelp reviews can be transformed into actionable insights for restaurant owners. By applying text analytics, the aim is to uncover what drives customer satisfaction and dissatisfaction. This analysis focuses on identifying patterns and trends in star ratings, pricing categories, and popularity rankings, ultimately providing recommendations for improving customer experience and boosting business performance. This report is structured to cover data preparation, visual insights from EDA, and a sentiment-based understanding of review texts.

Data Exploration

The dataset contains **2,381 observations across 10 variables**, including features such as star rating, price range, restaurant style, location, and customer reviews. Our focus during data exploration was to clean the dataset for meaningful analysis:

- **Price Category Handling**:
  The price variable is categorical and consists of four distinct levels: $(cheapest), $$, $$$, and $$$$(costliest). However, a significant number of entries had missing values. Rather than dropping those records, the missing values were imputed using the **mode** of the most common price category.

- **Star Ratings**:
  The star rating represents the **average rating of the customer reviews** and is a direct measure of satisfaction. However, without the context of what customers said, it's hard to understand *why* a restaurant got a specific rating. That's where sentiment and keyword analysis of reviews become crucial to get a deeper understanding.

- **Restaurant Style and Location**:
  The restaurant style column contains **134 unique values**, ranging from cuisine types to service styles. Due to this high cardinality, it was impractical to categorize them within the scope of this project. Similarly, location is known to affect restaurant success, but this study keeps its scope limited to **online feedback metrics** to ensure outcomes focused on customer satisfaction.

Methodology

Text Pre-Processing

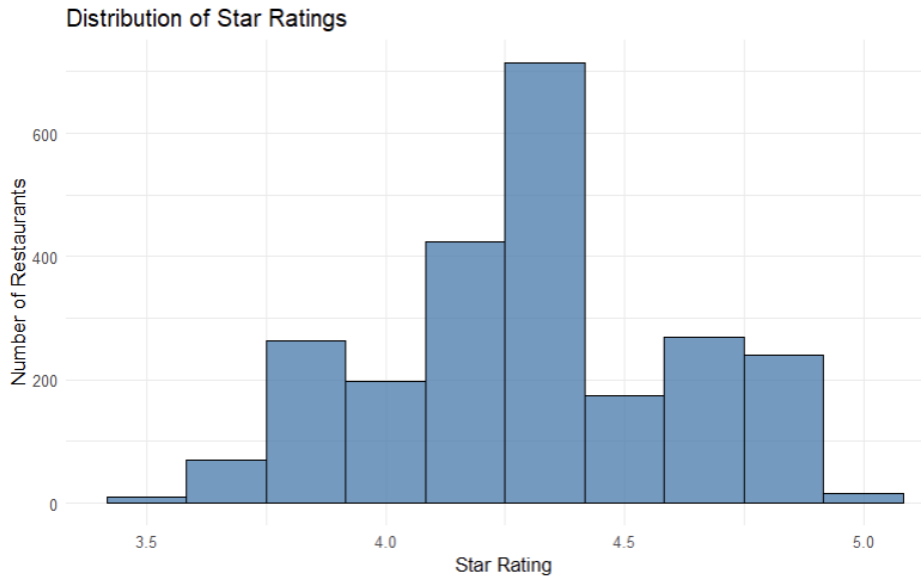
The main focus was on cleaning the **review comments** column to prepare it for text analysis. Removed special characters, numbers, punctuation, and common English stop words, converted all text to lowercase, and tokenized the text into individual words. Finally, **lemmatization** was applied to reduce words to their base form (e.g., "serving" → "serve")

Explanatory Data Analysis

An EDA was conducted to uncover patterns in star ratings, pricing, and popularity across restaurants. Visualizations helped highlight relationships that boost customer satisfaction and business performance.

1. Distribution of Star Ratings

This histogram shows the distribution of average star ratings across restaurants. Most ratings fall between **4.0 and 4.5**, indicating a generally favorable customer sentiment. A peak around **4.3** suggests many restaurants maintain good but not perfect ratings. Very few restaurants receive ratings below 3.5 or above 4.8, which implies that extreme satisfaction or dissatisfaction is rare. This distribution helps identify the baseline satisfaction level and where most restaurants cluster.



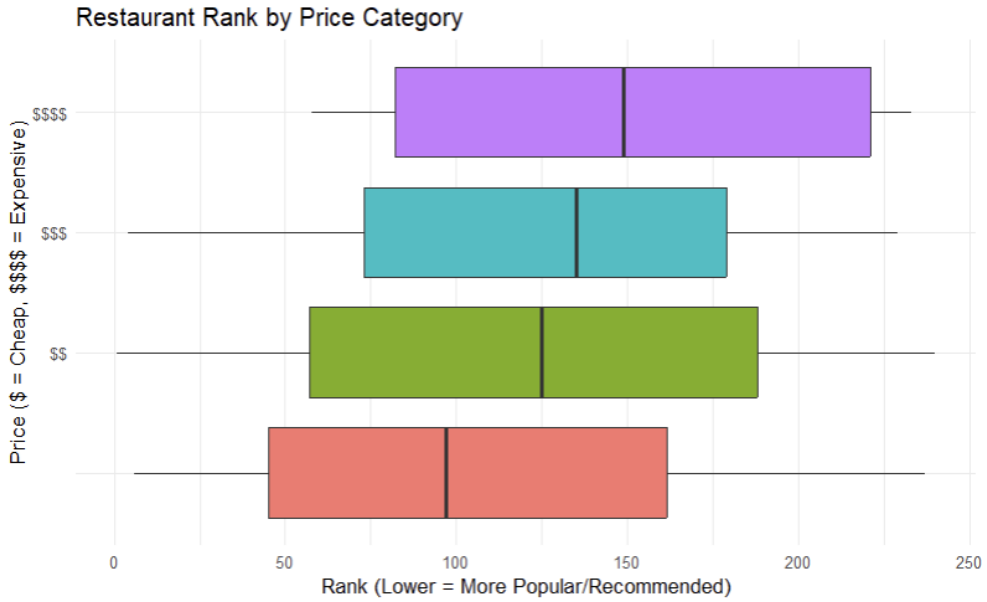
2. Restaurant Rank by Price Category

This box plot explores how restaurant ranking varies across different price categories:

1. **Cheaper restaurants ($)** show a wider range in popularity, suggesting that affordability doesn't compromise popularity.

2. **Mid-range options ($$ and $$$)** have a more balanced rank distribution, indicating steady performance and consistent appeal.
3. **Expensive restaurants ($$$$)** tend to have higher median ranks (less popular), though some still perform very well, indicating niche popularity or premium positioning.

These findings may be surprising on the surface. From this, one can infer that while expensive restaurants may not always be top-ranked, affordability combined with quality can drive popularity.



Word Clouds

To get a sense of what customers talk about most in their reviews, a word cloud was created using all the feedback. As expected, common terms like "good," "food," "place," "service," and "order" stood out clearly. While these are typical in restaurant reviews, they don't offer much insight on their own. However, specific words such as "flavor," "dish," and "ambiance" suggest that customers often comment on the taste of food and the overall dining atmosphere.

Comparative Word Clouds by Star Rating

A detailed analysis was conducted of word clouds with 3-star, 4-star, and 5-star ratings to uncover what differentiates restaurants at each level of customer satisfaction. The tone, context, and additional descriptors surrounding these words vary with star level, offering an idea of what drives positive reviews.

3-Star Ratings

The 3-star word cloud reveals a mix of neutral and slightly critical feedback. Presence of logistical and experience-related terms such as *parking*, *reservation*, *busy*, and *wait* suggests that diners found the experience adequate but not memorable, perhaps hindered by inefficiencies or unmet expectations. Positive terms like *recommend* and *friendly* do appear, but with less enthusiasm compared to higher ratings. The frequent mention of *server* and *ambiance* implies that the experience may have lacked refinement.

4-Star Ratings

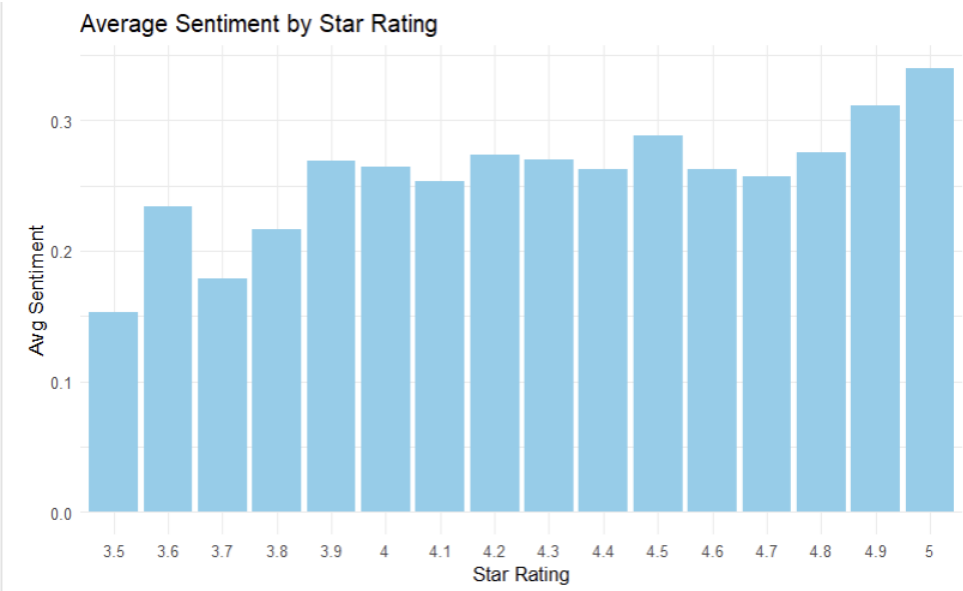
The 4-star reviews reflect a more favorable dining experience. Words like *amazing*, *delicious*, *favorite*, begin to dominate the feedback, suggesting strong appreciation for the food quality and overall experience. Though terms like *wait* and *parking* still make appearances, they are overshadowed by compliments related to the dishes and service. Specific food items such as *pork*, *salad* emerge, indicating engagement with the menu and dining experience. There is also a greater emphasis on *ambiance* and *love*,

showing that customers feel more emotionally satisfied and are more likely to return or recommend the place.

5-Star Ratings

The 5-star word cloud stands out with words such as *amazing*, *perfect*, *loved*, *friendly*, and *attentive* dominating the reviews, indicating that customers felt genuinely delighted and valued. There is also a noticeable increase in specific and culturally rich food descriptors like *korean*, *kimchi*, *spicy*, *fried*, *beef*, and *fresh*, pointing to an appreciation for authenticity, taste diversity, and culinary excellence. The presence of *service*, *flavor*, and *experience* reinforces that both the food and the overall atmosphere contributed equally to customer satisfaction. This level of detail suggests that customers were not only pleased but also emotionally connected to their experience.

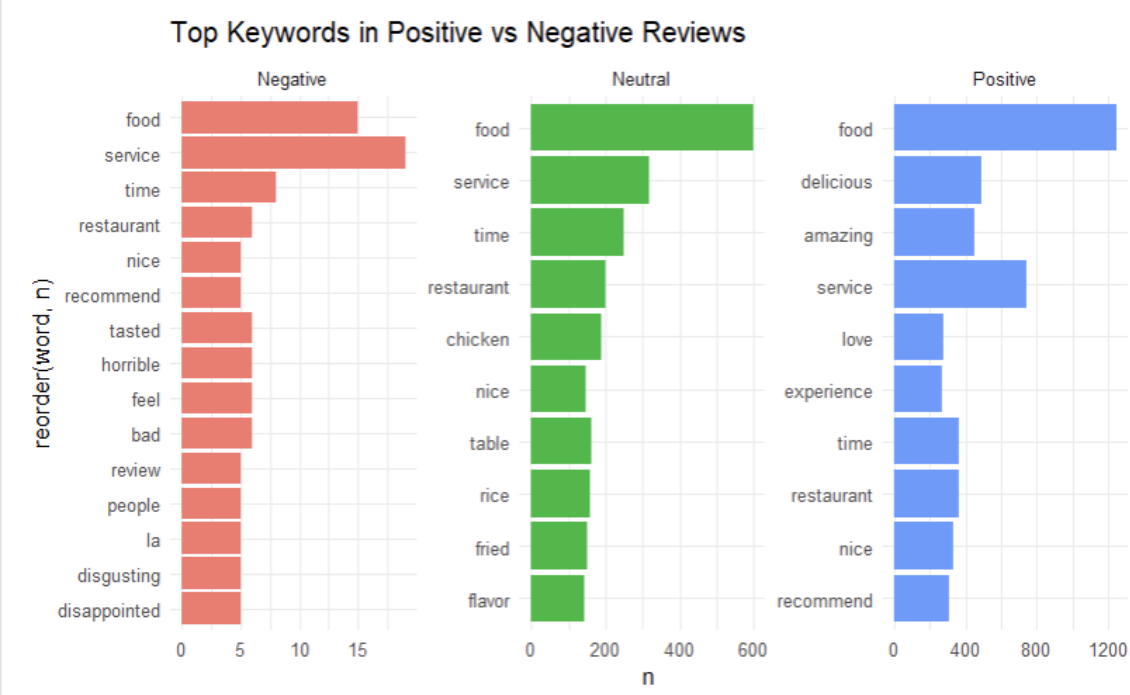
Sentiment Analysis



The bar chart highlights customer satisfaction by combining star ratings with review sentiment. While star ratings provide a numerical score, sentiment analysis reveals the emotional tone behind it. A clear **positive correlation** is observed; lower ratings (3.5–3.7) show lower sentiment scores (0.15–0.2), indicating more negative or mixed feelings. In contrast, higher ratings (4.0 and above) align with stronger positive sentiment, peaking above 0.3 for 5-star reviews.

Top Keywords in Positive vs Negative Reviews

This tripartite bar chart breaks down the most common keywords used in **negative**, **neutral**, and **positive** reviews. Unsurprisingly, the words *"food"* and *"service"* dominate across all sentiment categories. However, the contexts differ as while *"food"* appears alongside words like *"bad"*, and *"disappointed"* in negative reviews, in positive reviews it co-occurs with uplifting words like *"delicious"*, *"amazing"*, "love". . This shows that both the food and the staff experience are critical touchpoints.

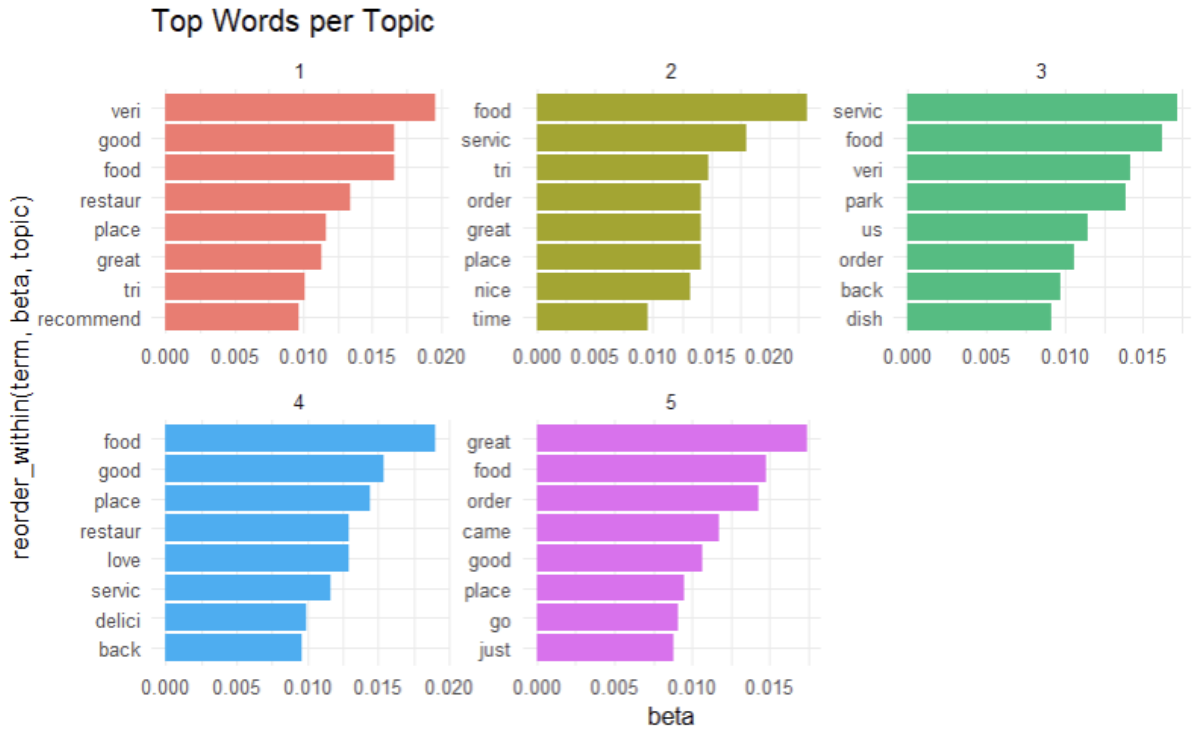


Top Words per Topic (Topic Modeling)

The five panels here display the **top words per topic** extracted using a topic modeling technique such as LDA (Latent Dirichlet Allocation).

- **Topic 1** focuses on general satisfaction with words: *"very"*, *"good"*, *"recommend"*, *"place"*
- **Topic 2** blends service and order experience, with terms: *"try"*, *"order"*, and *"great"*
- **Topic 3** focuses on logistics and ambiance with words: *"park"*, *"back"*, and *"dish"*
- **Topic 4** emphasizes emotion and return intention: *"love"*, *"back"*, *"delicious"*
- **Topic 5** appears action-oriented, including terms:*"came"*, *"go"*, and *"just":* reflecting efficiency of service.

This breakdown is useful for understanding the **latent themes** in the data offering restaurants a lens on customer feedback whether it's about food quality, atmosphere, or service efficiency.



Predictive Modelling

Regression Model and KNN Performance

Two models were used for predictive modelling: **Linear Regression** and **K-Nearest Neighbors (KNN)** to predict customer ratings based on review text features.

1. The **Regression model** performance:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 26  2
##          1 13  1
##
##                Accuracy : 0.6429
##                  95% CI : (0.4803, 0.7845)
##     No Information Rate : 0.9286
##     P-Value [Acc > NIR] : 1.000000
##
```

2. The **KNN model** performance:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 39  3
##          1  0  0
##
##                Accuracy : 0.9286
##                  95% CI : (0.8052, 0.985)
##     No Information Rate : 0.9286
##     P-Value [Acc > NIR] : 0.6473
##
```

- The models Logistic Regression and K-Nearest Neighbors (KNN) predicted whether a review represented a perfect 5-star rating (class 1) or not (class 0), using TF-IDF features.
- Despite using SMOTE to address class imbalance in the training data, both models struggled to generalize effectively on the test set, which remained heavily skewed toward class 0 (over 92%).
- The Logistic Regression model offered moderate classification performance, while the KNN model achieved high accuracy but failed to detect any class 1 instances, revealing a lack of sensitivity to minority cases.
- These outcomes highlight the challenges of working with imbalanced data and suggest that further tuning or alternative modeling strategies may be required for better performance.

Business Recommendations

Food and service are core across all ratings, regardless of whether a review is 3-star, 4-star, or 5-star restaurant.

1. **Invest in Service Training and Staffing**
3-Star restaurants should strive to improve their logistics and focus on service, attentiveness, friendliness, and proactive support to turn average experiences into exceptional ones. They can streamline logistical issues like long wait times, limited parking, and reservation handling, which would elevate a customer's perception.

2. **Craft Memorable, Culturally Distinct Menus**
4-Star restaurants should focus on food enjoyment and moderate praise. Encourage staff to personalize service, celebrate special occasions, and create an aesthetic ambiance to spark joy. Offering authentic standout items can make the restaurant more memorable and boost reviews.

3. **Enhance the Ambiance and Visual Aesthetic**
5-Star restaurant reviews are emotionally rich and specific; fresh decor, appropriate music, and the vibrant atmosphere itself is an experience. To keep the five-star status, they could offer exclusive menu previews, VIP reservations, or surprise discounts to returning guests.

4. Monitor Sentiment Trends Over Time
Use sentiment analysis dashboards to track fluctuations in customer mood. A dip in sentiment for a specific period could flag an operational or menu issue before star ratings drop. Encourage detailed reviews; the more specific a review is, the higher the likelihood of it becoming a 5-star.

5. **Use Review Text Alongside Other Metrics**
While textual analysis is insightful, combining it with pricing, location, and competitor comparisons can lead to more accurate predictive modeling and operational strategy.

Conclusion

Our project extracted meaningful insights from restaurant reviews using sentiment analysis, topic modeling, clustering, and predictive modeling. It was observed that emotionally rich and food-focused language correlated with higher ratings, while lower ratings centered around service and value-related terms. In predictive modeling, both Logistic Regression and KNN showed potential but struggled due to class imbalance, especially in detecting rare 5-star reviews. These insights can serve as valuable guidelines for struggling restaurants seeking to improve their reputation, as well as for new establishments entering the competitive food industry.