Data Mining Project

Angelina Benyahia
12/2024

Part I: Introduction

A. Dataset

The dataset chosen for this project is called *Recipe Reviews and User feedback,* a csv file containing 18182 instances, and 15 features, which responds to the criteria imposed. Classification and Regression can be applied to it.

The subject touches on feedback certain recipes receive: the data indicates a recipes' ranking on the top 100 recipes list (recipe_number), users' sentiment towards recipes quantified on a 1 to 5 star rating scale, with a score of 0 denoting an absence of rating, as well as textual reviews. Each user has an internal user reputation score . Each review comment is uniquely identified with a comment ID and comes with additional attributes, including the creation timestamp, reply count, and the number of up-votes and down-votes received (thumbs up, thumbs down).

This dataset is a valuable resource for researchers and data scientists, facilitating endeavors in sentiment analysis, user behavior analysis, recipe recommendation systems, and more. It offers a window into the dynamics of recipe reviews and user feedback within the culinary website domain.

B. Objectives

Through classification, we want to group the recipes into 2 categories: good recipes, and bad recipes. Through regression, we want to predict the sentiment or rating of customer feedback based on the recipe review text. Additionally, we will identify patterns in recipe names or feedback that correlate with higher ratings.

Part II. Descriptive Analysis of the Dataset:

A. Cleaning the data:

Before training the data, it is important to clean the dataset to avoid bias and errors. Firstly, we detect and correct the missing values (2 in text, replaced by ‘unknown’), and identify duplicates. We prepare the data numerically: we can drop comment_id and user_id, as these are redundant with the user_name. The non-numerical values left are user_name and recipe_name, and the text reviews, which we need to keep as is.

The text needs to be treated individually, with the NLTK library: To clean text for analysis, we start by converting it to lowercase for consistency, removing punctuation, and extra whitespace. We eliminate stop words (e.g., "the," "and") to focus on meaningful terms, and tokenize the text into individual words. Optionally, handle special characters, emojis, or spelling corrections. Finally, we remove any domain-specific noise (like ‘recipe’ ‘made’) and rejoin the cleaned tokens.

B. Feature Engineering:

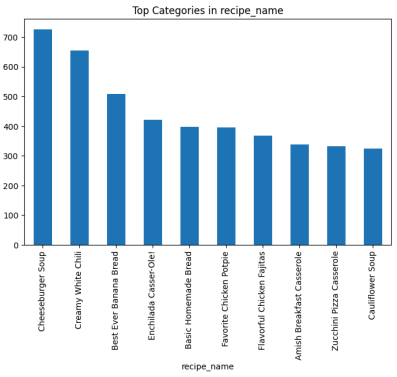
In this feature engineering process, we heavily focused on transforming the textual reviews to more exploitable data. Several new columns were created to enhance the dataset. First, we identified the presence of "good" and "bad" keywords in recipe texts, adding columns for keyword counts and binary indicators of their presence. Next, sentiment analysis was performed using TextBlob to generate sentiment scores, which were rescaled to align with the star rating scale. Aggregated features, such as the average star ratings and rescaled sentiment scores, were computed for each recipe. Finally, a weighted grade was calculated by combining the average star ratings (weighted at 60%) and the rescaled sentiment scores (weighted at 40%), allowing for a more nuanced evaluation of each recipe.

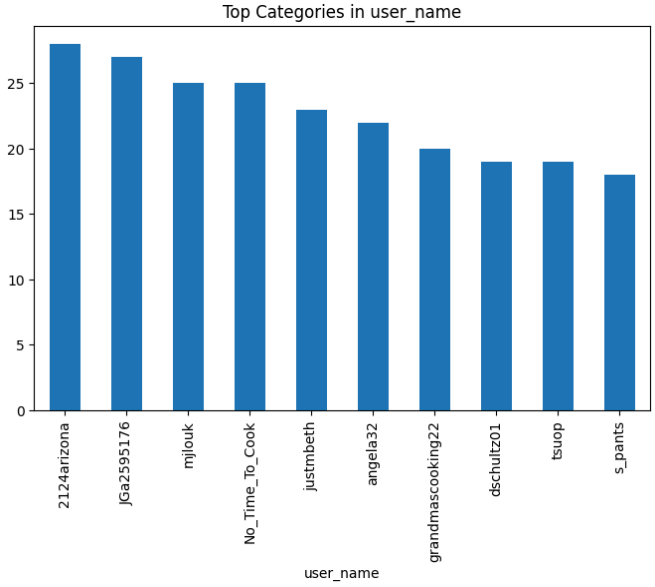
These are all the columns after feature engineering:
'Unnamed: 0', 'recipe_number', 'recipe_code', 'recipe_name', 'user_name', 'user_reputation', 'reply_count', 'thumbs_up', 'thumbs_down', 'stars', 'best_score', 'text', 'processed_reviews', 'good_keyword_count', 'bad_keyword_count', 'has_good_keywords', 'has_bad_keywords', 'sentiment_score', 'sentiment', 'rescaled_sentiment_score'

C. Univariate Analysis

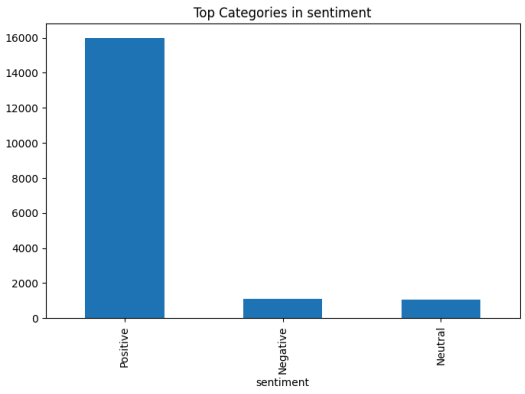
These outputs will provide more detail on our dataset features.

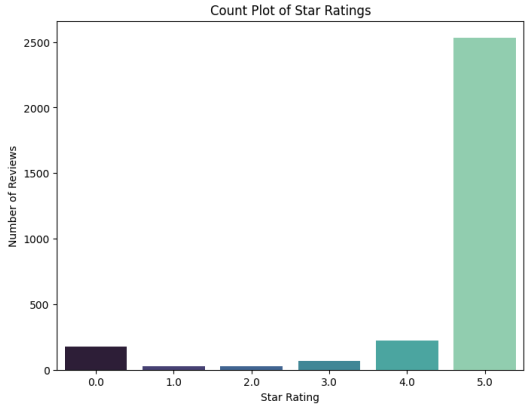
The most common Recipes:		The least common recipes:	
Cheeseburger Soup	725	Peanut Butter Cup Cheesecake	96
Creamy White Chili	654	Blueberry French Toast	90
Best Ever Banana Bread	509	Caramel Heavenlies	86
Enchilada Casser-Ole!	421	Lime Chicken Tacos	86
Basic Homemade Bread	397	Vegetarian Linguine	31





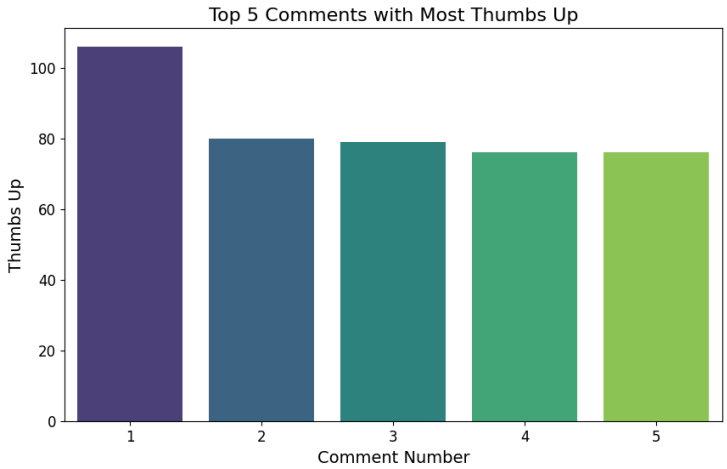
The most active users are 2124arizona, JGa2595176, and mjlouk, with over 25 reviews each. As these users are more experienced, their input might be more valuable.





Most
recipes are well rated, with 5 stars out of 5. Those with a rating of 0 are the recipes that haven't received any . It will be interesting to study the recipes with low stars (1-3).

Top 5 Comments with the Most Thumbs Up:



1. Username: C `Recipe : Skillet Shepherd's Pie`
Comment: Good,easy recipe. For all u snarky bullies who feel the need to comment that it's not technically Shepherds Pie - please get over yourselves. Many recipes that refer to themselves as spaghetti sauce or enchiladas are not necessarily authentic, but rather americanized versions of delicious dishes. Smh
Thumbs Up: 106, Thumbs Down: 3

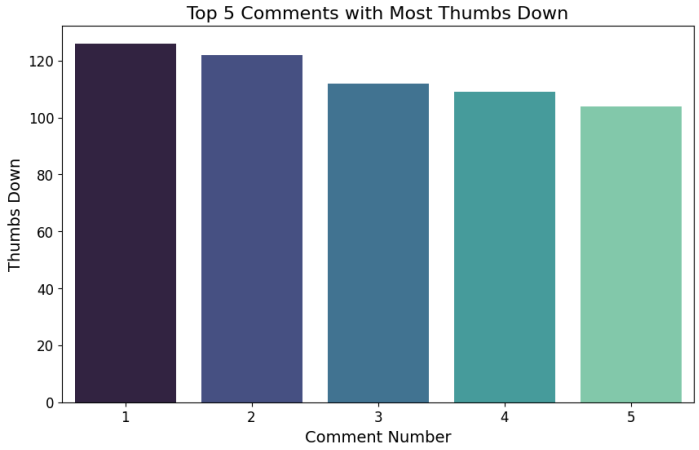
2. Username: Murray111 `Recipe : Contest-Winning New England Clam Chowder`
Comment: Fat Free Half and Half is not Half and Half. It is skim milk and a lot of sugar. Heavy Cream or real full fat half and half are far better than the chemical storm known as fat free half and half. It actually should not be legally called half and half its skim milk and sugar. Devoid of flavor and substance filled with nasty chemicals and sugar. Why do we have an obesity problem in this country because of misnomers like this. dictionary says a wrong or inaccurate use of a name or term when speaking of fat free half and half.
Thumbs Up: 80, Thumbs Down: 14

3. Username: JulianneMasterson `Recipe : Pork Chops with Scalloped Potatoes`
Comment: I enjoys these reviews. Rarely make the recipes. Jaw-dropping amazing how mean some of these reviews are. That someone would sign in and go out of their way to tell someone "learn to cook", " this was terrible", what's wrong with you people, this is a recipe site! Why do you have to be infect your personality onto this nice place by being jerks?
Thumbs Up: 79, Thumbs Down: 5

4. Username: karen794 `Recipe : First-Place Coconut Macaroons`
Comment: I like the ingredients and portions, but this is how I put them together. In an upright mixer, beat the egg whites until they are stiff peaks, add vanilla. In another bowl, mix the other ingredients and fold them into the egg whites.
I baked them on parchment paper and they were delicious
Thumbs Up: 76, Thumbs Down: 18

5. Username: Kim5287 `Recipe : Porcupine Meatballs`
Comment: I remember a time long long ago lol..when my mom was making this for our family. I was a teenager and my little brother was about 5. I remember thinking it looked so good. Mom was taking toothpicks and stabbing them ..giving us all one to sample. Just as I was about to sample what I though looked and smelled fabulous my little brother screamed at my mom, stomped his foot and yelled No way! No way I am not eating stinky. I about died. I laughed so hard I cried. From that day forward me nor my brother would eat mamas porcupine meatballs. Tomorrow is my brothers 47th birthday. I am skipping the

cake and bringing stinky. I can't wait to see his face. Happy Birthday Mikey. This also dispelled that old rumor that Mikey will eat anything . he won't. Thumbs Up: 76, Thumbs Down: 9



Top 5 Comments with the Most Thumbs Down:

1. Username: Landon `Recipe : Simple Taco Soup`
 Comment: I love it it is the BEST SOUP EVER that is why i gave it one star
  Thumbs Up: 5, Thumbs Down: 126

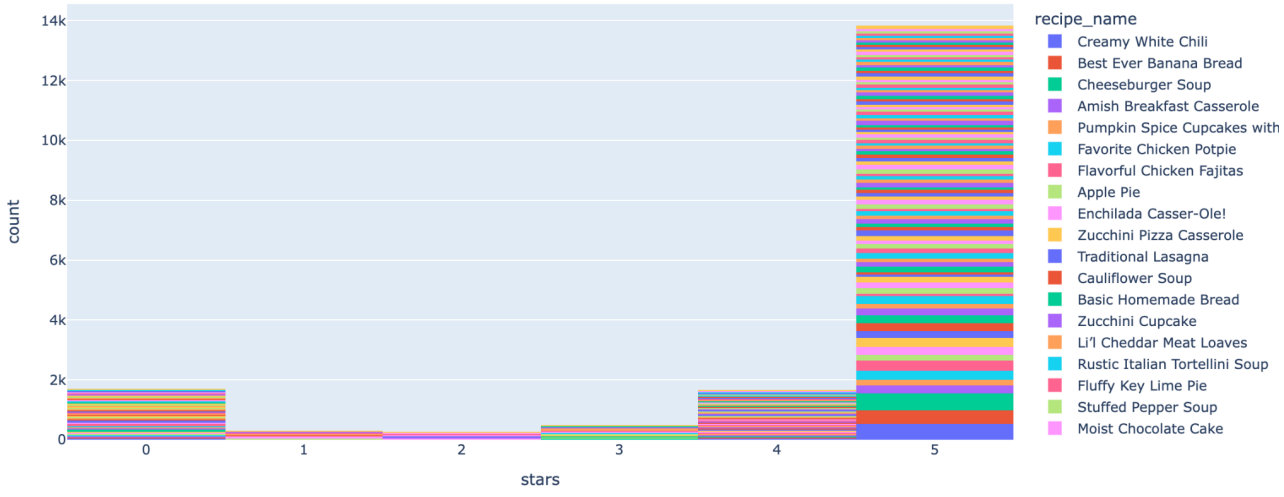
2. Username: SylCan `Recipe : Taco Lasagna`
 Comment: Maybe Im a little prejudiced but I have trouble trusting a recipe that uses a spice MIX. What's up with that? How difficut is it to shake out the cumin and corriander?
  Thumbs Up: 1, Thumbs Down: 122

3. Username: lizofstad231 `Recipe : Creamy Grape Salad`
 Comment: Nasty. Doesn'tt taste good unless you put tons of brown sugar on it. Otherwise, it tastes likes grapes with mayonnaise. Gross. My friends hated it. What a waste of money.
  Thumbs Up: 6, Thumbs Down: 112

4. Username: Ala406 `Recipe : Traditional Lasagna`
 Comment: Don't call it traditional and include cottage cheese. What a disaster.
  Thumbs Up: 48, Thumbs Down: 109

5. Username: Dale732 `Recipe : Skillet Shepherd's Pie`
Comment: I am always amazed at the lack of knowledge of British dishes.
  Thumbs Up: 13, Thumbs Down: 104

Recipes classed by star rating



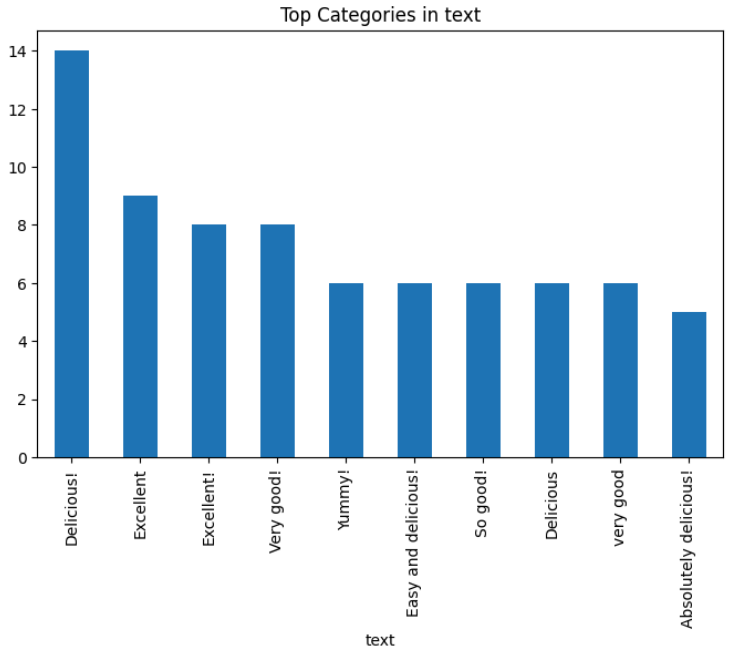
Basic Statistics for Stars Rating:

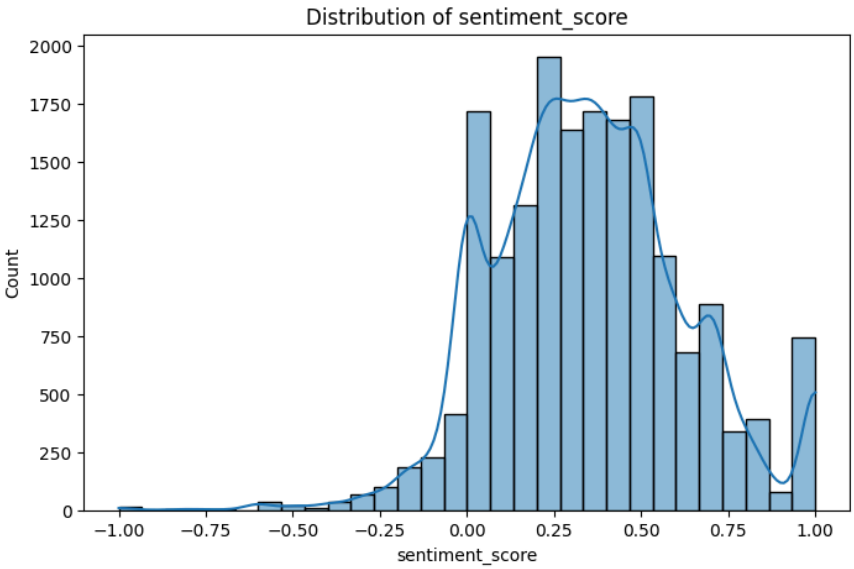
```
Stars column stats (excluding zeros):
count    16486.000000
mean         4.730013
std          0.738116
min          1.000000
25%          5.000000
50%          5.000000
75%          5.000000
max          5.000000
Name: stars, dtype: float64
Median of Stars (excluding zeros): 5.0
Min Stars: 1 (we exclude 0 as it is just an absence of rating)
Max Stars: 5
```

C.2 Univariate textual analysis

This word cloud displays the most used words amongst the reviews.

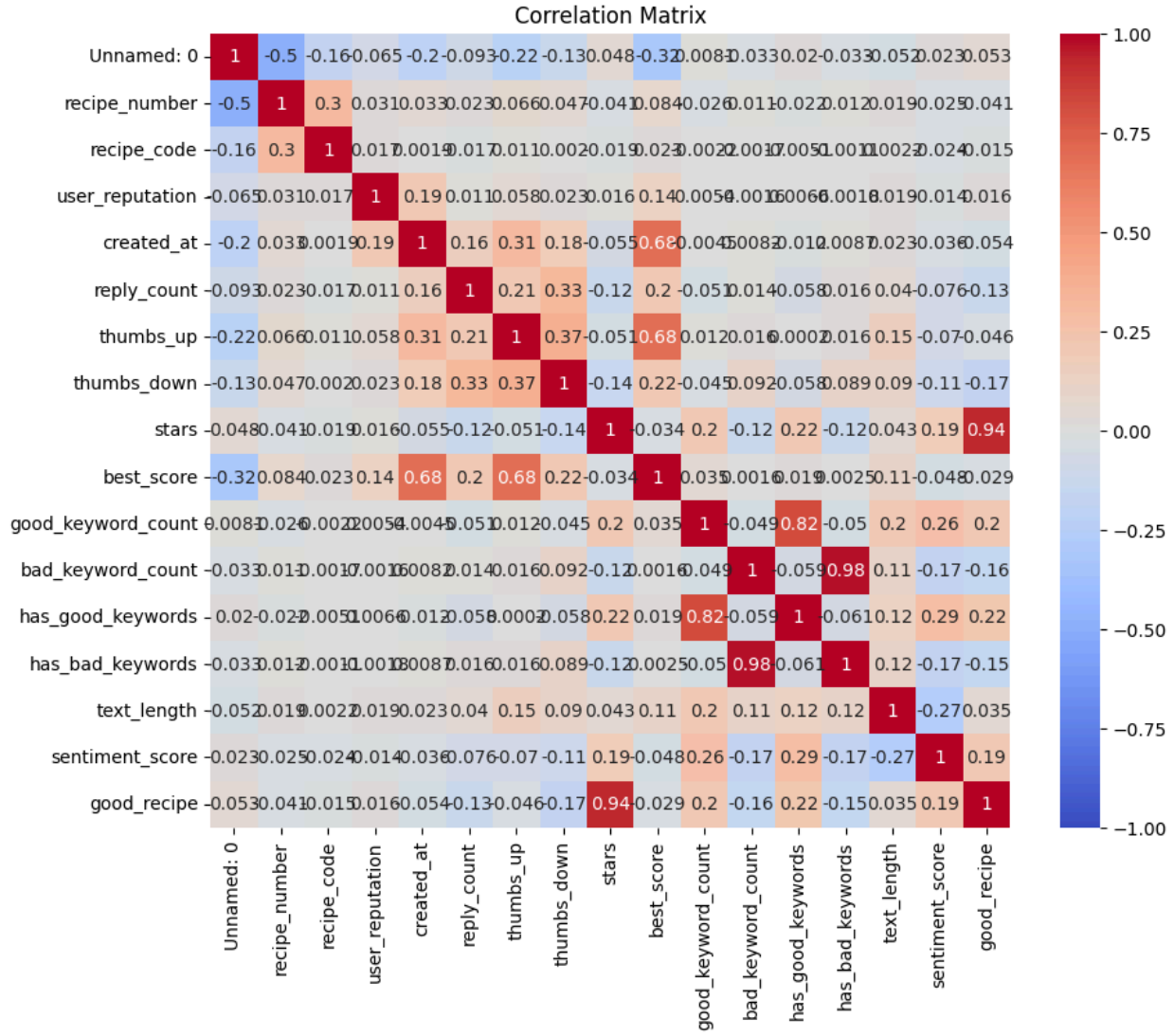
```
[('time', 3998), ('good', 3580), ('delicious', 3049), ('great', 2965),
('added', 2795), ('easy', 2789), ('love', 2653), ('one', 2616), ('family',
2479), ('like', 2463)]
```





Most comments or reviews in the dataset might lean toward neutral or positive sentiment (clustered between 0 and 0.5).
The lower frequency of negative scores suggests that negative sentiments are relatively rare in the data.

D. Bivariate statistics:



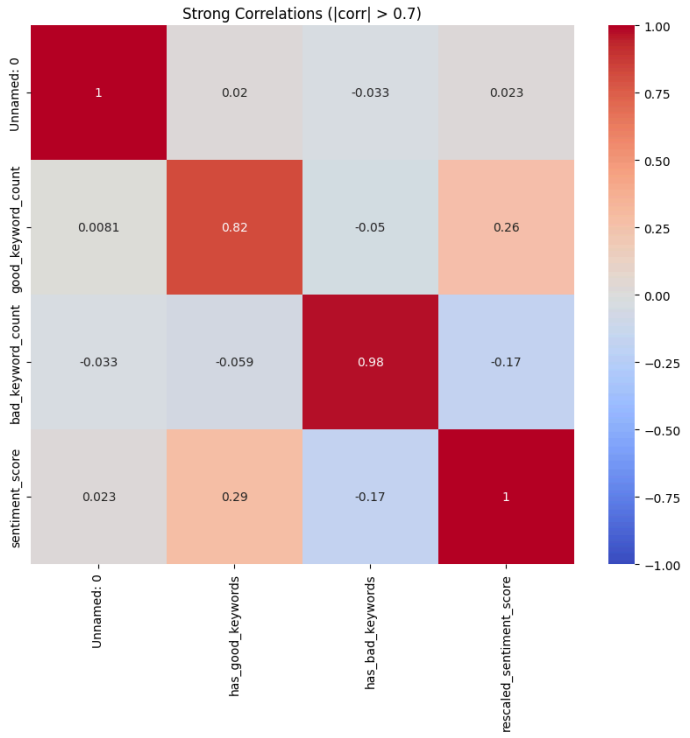
The correlation matrix highlights key relationships in the dataset.

- Stars and good_recipe show a very strong positive correlation (0.94), indicating highly rated recipes are often classified as good. Similarly, best_score correlates well with stars (0.68). Strong links are seen between good_keyword_count and has_good_keywords (0.82).
- Moderate correlations exist between thumbs_up and thumbs_down (0.33), while variables like text_length and sentiment_score have weak correlations, suggesting independence.
- No strong negative correlations are present, indicating minimal inverse relationships.

Top correlations:

Strong correlations (above threshold of 0.7):

- stars - good_recipe 0.937187
- good_keyword_count - has_good_keywords 0.823804
- bad_keyword_count - has_bad_keywords 0.975084
- Sentiment - rescaled_sentiment_score 1.000000



The heatmap highlights strong correlations (absolute correlation > 0.7) among specific variables in the dataset.

- Stars and Good Recipe (0.937): The high correlation indicates that higher star ratings strongly align with recipes classified as "good." This validates the categorization process, as user ratings are a direct measure of recipe quality.
- Good Keyword Count and Has Good Keywords (0.824): This relationship suggests that recipes identified as containing positive keywords are likely to have higher counts of these keywords. It supports the hypothesis that keyword presence can predict recipe quality.
- Bad Keyword Count and Has Bad Keywords (0.975): The strong correlation emphasizes the impact of negative keywords in reviews. Recipes with a high count of these terms are reliably marked as having negative keywords, highlighting a robust feature for classification.
- Sentiment Metrics (1.000): The perfect correlation between sentiment score and its rescaled version confirms that the scaling process did not introduce discrepancies, ensuring consistency in sentiment analysis.

2) Part III: Machine Learning Techniques

We chose to apply classification and regression techniques to analyse this dataset, through a logistic regression and a linear regression. Classification is well-suited to categorizing recipes into "good" and "bad" based on user feedback, as this involves a clear outcome variable (binary labels) derived from existing ratings or sentiment scores. Regression, on the other hand, is ideal for predicting numerical values, such as customer sentiment or star ratings, based on textual or numerical features. In contrast, clustering is unsupervised and would group recipes based on inherent patterns, which might not directly align with our goal of understanding sentiment or feedback quality. By focusing on classification and regression, we leverage the labeled data and address actionable questions: identifying factors that make recipes more appealing and predicting future ratings, making the analysis both practical and insightful for decision-making.

A. Cassification Model

With this model, we aim to group recipes into good recipes and bad recipes., through a logistic regression. Prior to training the model, we did some feature engineering by adding dummy features for flagged "good" or "bad" keywords in the reviews, such as “good, delicious, loved” or “disgusting, gross, bad”. These keywords can act as additional signals to improve the classification model.

After training the doing a 80-20% train test split on the dataset to categorize recipes as good or bad, these are the results obtained:

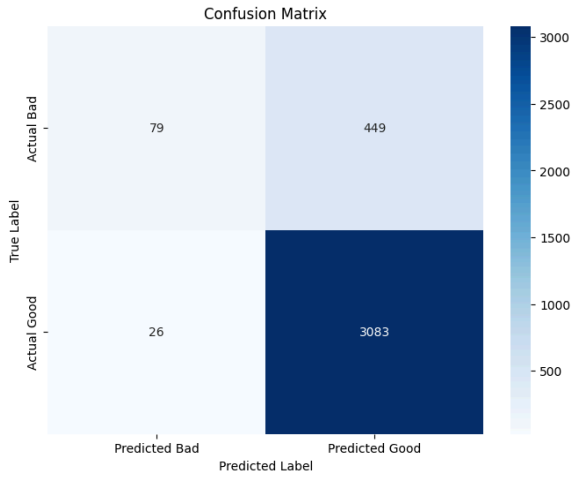
```
Classification Report:
               precision    recall  f1-score   support

           0       0.75      0.15      0.25       528
           1       0.87      0.99      0.93      3109

    accuracy                           0.87      3637
   macro avg       0.81      0.57      0.59      3637
weighted avg       0.85      0.87      0.83      3637

Accuracy Score: 0.8691229034918889
```

The model achieves strong overall accuracy (87%) but struggles with "bad recipes" (class 0), showing low recall (15%). It performs well in identifying "good recipes" (class 1) with high precision (87%) and recall (99%). This imbalance suggests the need to improve detection of "bad recipes," aligning better with our goal of accurately categorizing recipes for user insights. Balancing the dataset or adjusting thresholds could help refine performance.



The confusion matrix reveals the model's strong performance in predicting "good recipes," with 3,083 true positives and only 26 false negatives. However, it struggles with "bad recipes," correctly identifying only 78 out of 528 while misclassifying 450 as good recipes. This highlights the class imbalance and the difficulty in detecting less frequent "bad recipes."

Its complicated to fix, because of nuances in the text. English speakers make many false “positive comments” like not bad, or not good , not horrible, which can be hard to interpret correctly through sentiment analysis.

B. Regression Model - Linear Regression

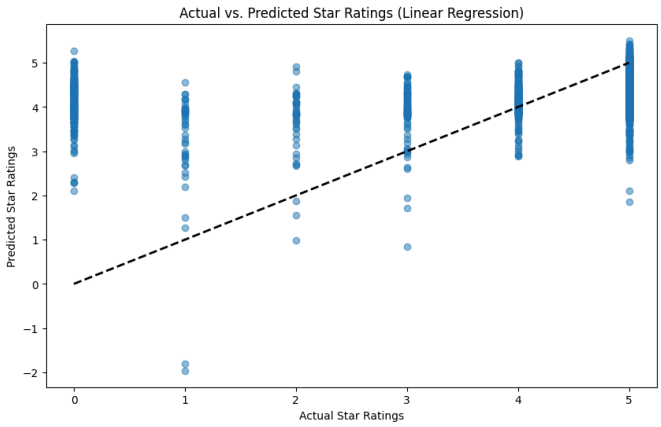
We want to use the linear regression model to predict the stars or sentiment based on review text and other relevant features.

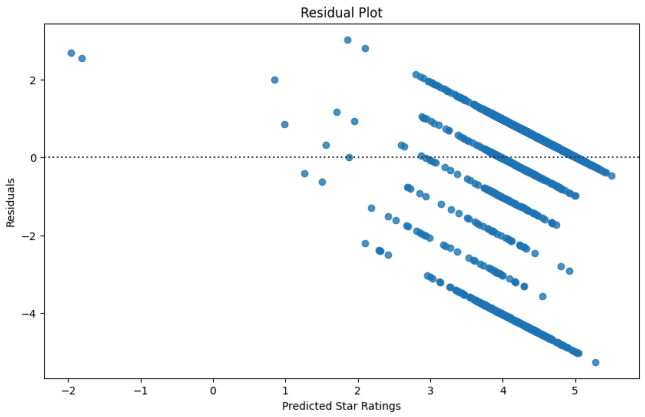
Results obtained:

```
R-squared: 0.07879137324052476

Mean Squared Error: 2.1734354115906767
```

R2 shows that only 7.88% of the variance in the stars (target variable) is explained by the features in the model, which is low. It suggests that the chosen features (reply_count, thumbs_up, thumbs_down, user_reputation, sentiment_score) are not strongly predictive of the recipe ratings. There may be missing key predictors or a non-linear relationship that the model cannot capture.
An MSE of 2.17 suggests that predictions deviate significantly from the actual ratings, which are on a 1-5 scale.
While it is hard to judge this value without a benchmark, the high MSE aligns with the low R-squared, indicating room for improvement in model performance.





The points are not aligned with the line, indicating that the model did not perform correctly.

Conclusion

This project explored the application of data mining techniques to analyze recipe reviews and user feedback, yielding valuable insights into user sentiment and recipe quality. Through descriptive analysis, we identified patterns in user interactions and recipe ratings, highlighting key trends and correlations. Feature engineering, including sentiment analysis and keyword extraction, proved critical for enhancing the dataset's predictive power.

The classification model effectively categorized recipes as "good" or "bad," achieving an accuracy of 87%. However, the imbalance in detecting "bad recipes" underscores the challenges posed by class imbalance and nuanced language in reviews. Future improvements could focus on advanced text-processing techniques or balancing strategies to refine model performance further.

The regression model, while less successful, revealed potential gaps in feature selection and highlighted the need for more sophisticated techniques or additional predictors to improve predictive accuracy. Despite its limitations, it provided valuable insights into sentiment drivers for recipe ratings.

Overall, this project demonstrated the power of combining data preprocessing, feature engineering, and machine learning to gain actionable insights from user-generated content. Further enhancements, such as deep learning for sentiment analysis or exploring non-linear models, could elevate the analysis and offer deeper understanding for stakeholders in the culinary domain.