# Data Appendix

We used two datasets from the Kaggle dataset (https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions). These two datasets are named RAW_ interactions and RAW_recipes.

## RAW_interactions

The dataset RAW_interactions contains 5 columns. The columns are called user_id, recipe_id, date, # rating, and review. Each row/entry represents a review and rating from a user on a Food.com recipe.

## RAW_recipes

The dataset RAW_recipes contains 12 columns. The columns are called name, id, # minutes, contributor_id, submitted, tags, nutrition, n_steps, steps, description, ingredients, and n_ingredients. Each row/entry represents a Food.com recipe.

## Our Final Dataset

After joining based on the recipe id (recipe_id and id), our final dataset that we used for our analysis contains 7 columns. The columns are called name, id, minutes, n_steps, rating, review, and sentiment_score. Each row represents a review and rating of a recipe.

### name

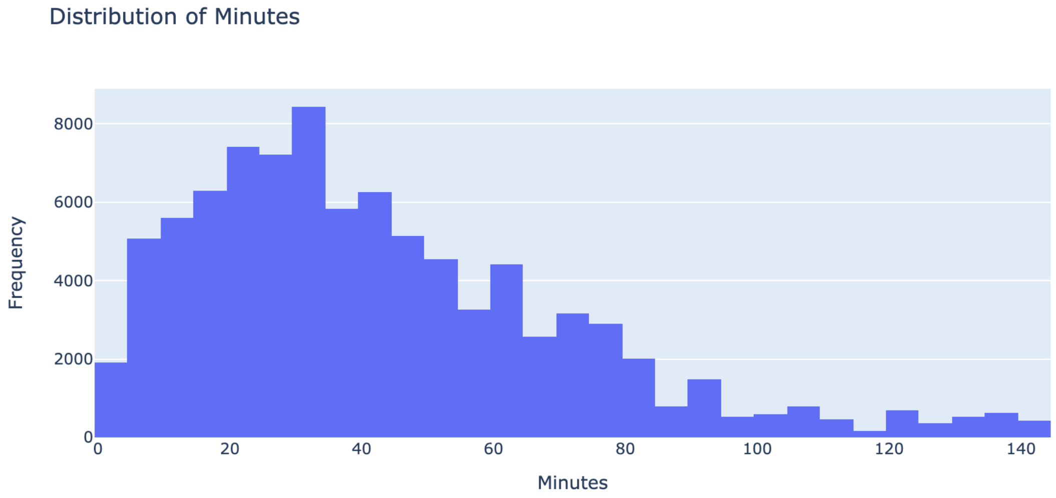
The name column is the submitted name of the recipe that has been reviewed. It is a string column. This column is not used in our analysis, but is included in our final dataset for clearer identification of individual recipes.

### recipe_id

The recipe_id column is a unique identifier of the recipe that is being reviewed and rated. It is a numeric column ranging from 2 to 6 digits in length. The recipe_id was used to join our datasets, but is not used in our analysis.

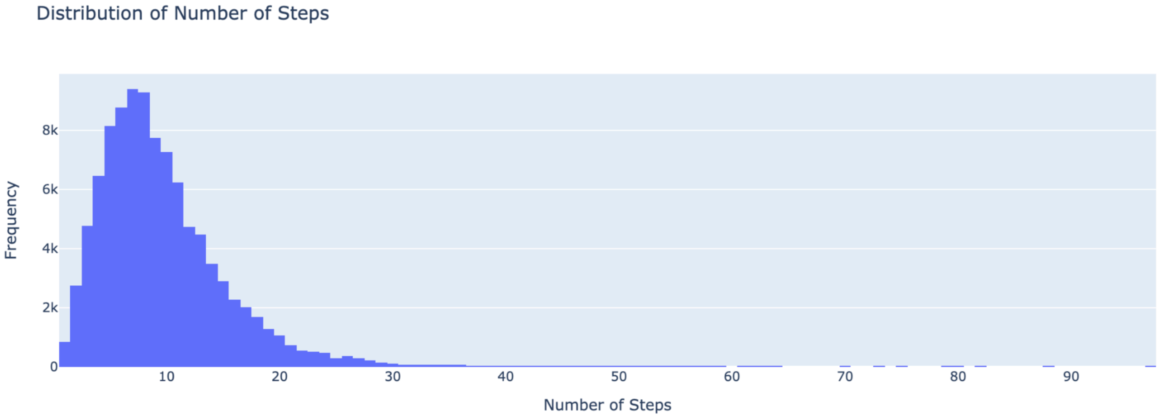
### minutes

The minutes column represents the estimated number of minutes it takes to complete the recipe from start to finish. It is a numeric variable. We are using this column in our analysis. The mode for minutes is around 35 minutes. The distribution of the minutes variable is skewed to the right. There is a smaller percentage of minutes greater than 100 minutes. Most are below 60 minutes.



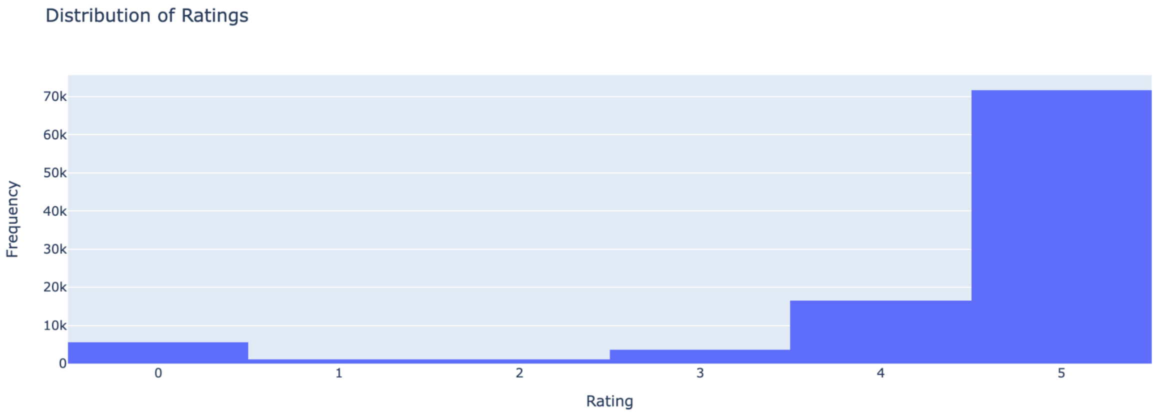
n_steps

The n_steps columns represent the number of steps in a given recipe. The number of steps is a positive integer value. We are using this column in our analysis. The n_steps column is right skewed as seen in the histogram below. There are lots of n_step values that look like outliers, above around 60 steps. The large majority of recipes have below 20 steps.



rating

The rating column represents the rating value given by a user. Ratings are integer values from 0 to 5, 0 representing the poorest rating and 5 being the highest rating. We are using this column for our analysis. The distribution of ratings is left skewed. There are only 6 possible values, and the most common choice is a rating of 5. There are a small number of 1 and 2 ratings in comparison to the rest.

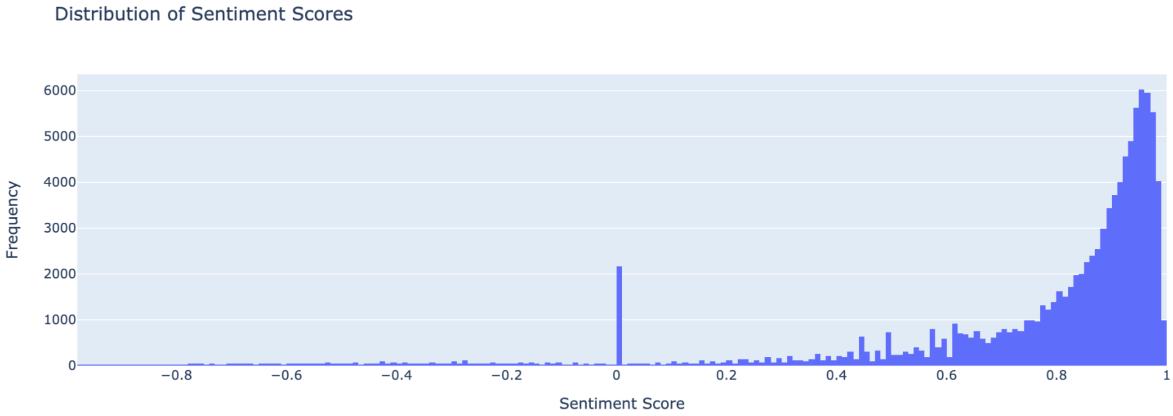


review

The review column is the review given by a user (user_id) for any given recipe (recipe_id). The entries text which are in a string format. We are using this column for our analysis. This is the column we are performing sentiment analysis on.

sentiment_score

The sentiment_score column is the numerical score calculated by VADER Sentiment Intensity Analyzer. It has a range from -1 to 1 and (-0.05, 0.05) is considered neutral. We are using this column for our analysis. This column was created by using VADER on the review column. The sentiment_score column is left skewed. Most of the sentiment score values are above around 0.6, which makes sense because most of the ratings are 5 out of 5. There is a spike in sentiment scores around 0, and then not very many negative sentiment scores past that.


Distribution of Sentiment Scores
Frequency
Sentiment Score