Recipe Recommender Assignment EDA

By Meraj Alam and Neha Netkar

OBJECTIVE OF THIS ASSIGNMENT

If you visit any recipe websites, let's say food.com, for example, you will notice a section called "You'll also love." Under this section, the website recommends recipes related to the one you are looking at or based on your past rating patterns. Our job is to design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at.

In this assignment we have to use AWS EMR to work on BigData to perform Exploratory Data Analysis and feature extraction from the raw data.

❖ Task 1

1. Read the Data:

Read the recipe data Read RAW_recipes.csv from the S3 bucket and ensured that each field had the correct data type.

❖ Task 2

2. Extract Individual Features:

One of the tasks involved splitting the nutrition column into seven individual columns to better understand the nutritional values. Separate the array into seven individual columns to create new columns named calories, total_fat_PDV, sugar_PDV, sodium_PDV, protein_PDV, saturated_fat_PDV, and carbohydrates_PDV.

❖ Task 3

3. Standardize Nutrition Values:

I standardized the nutritional values to per 100 calories to make comparisons fair across different recipes.

❖ Task 4

4. Convert Tags Column:

I converted the tags column from a string to an array of strings for easier analysis.

❖ Task 5

5. Join Second Data File:

Read the RAW_interaction.csv and join this interaction level file with the recipe level data frame. The resulting data frame should have all the interactions.

❖ Task 6

6. Create Time-Based Features:

Create features that capture the time passed between one review and the date on which the recipe was submitted. Use the review_date and the submitted columns after you join the two data files.

❖ Data Manipulation

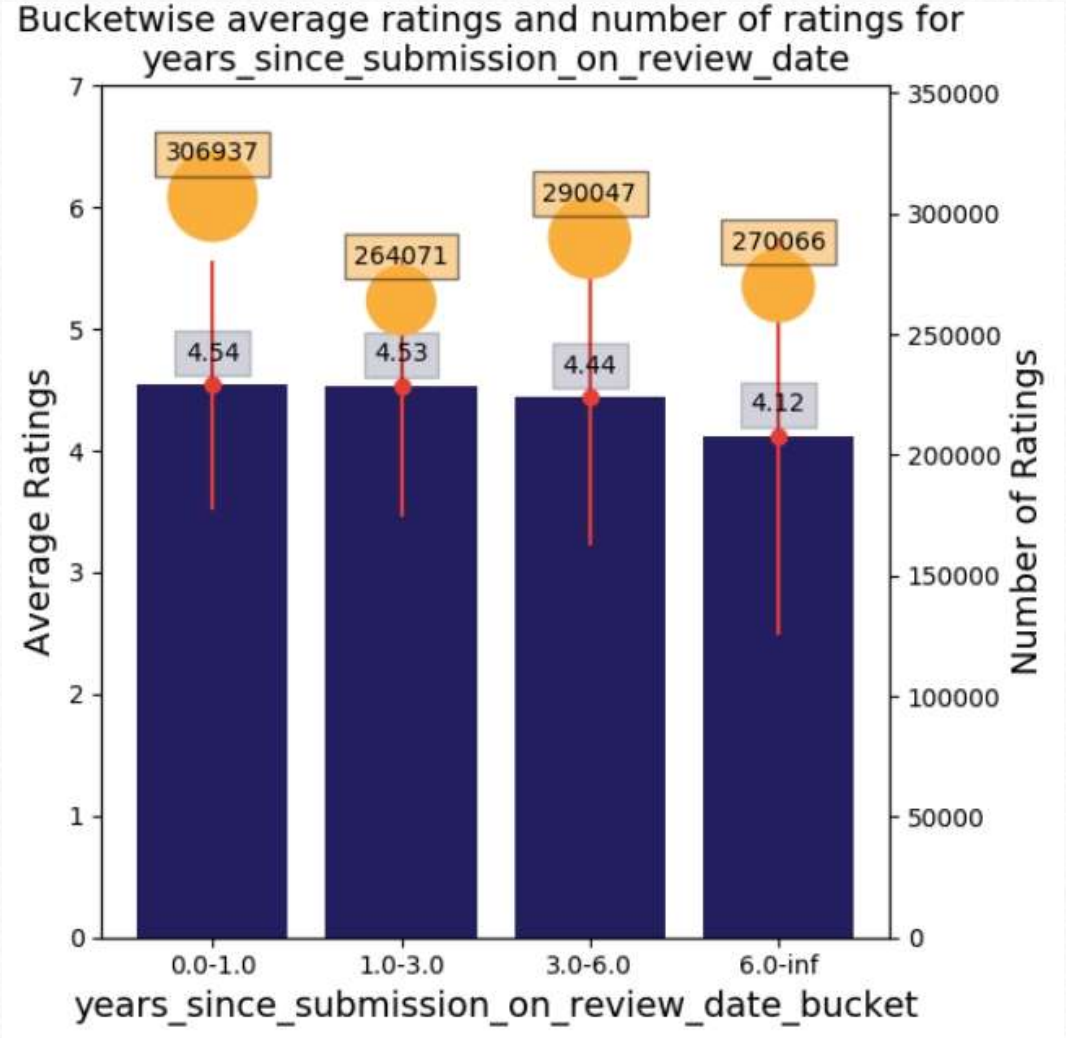
- Standardize the nutrition values
- By converting the nutrition values from absolute to relative terms, we ensure that portion size is not a factor in the analysis.
- All nutrition columns standardized to per 100 caloriesWe have included some test cases given below. You can use them to check if you have completed the task correctly
- Complete the code in the following cell
- Convert the tags column from a string to an array of strings
- Join Recipe Data to Review Data and Read the second data file
- Create time-based features
- Save the data we have created so far in a parquet file. ('s3a://upgradfoodrecsysdir/interaction_level_df_processed.parquet')

EDA

Years since submission on review date

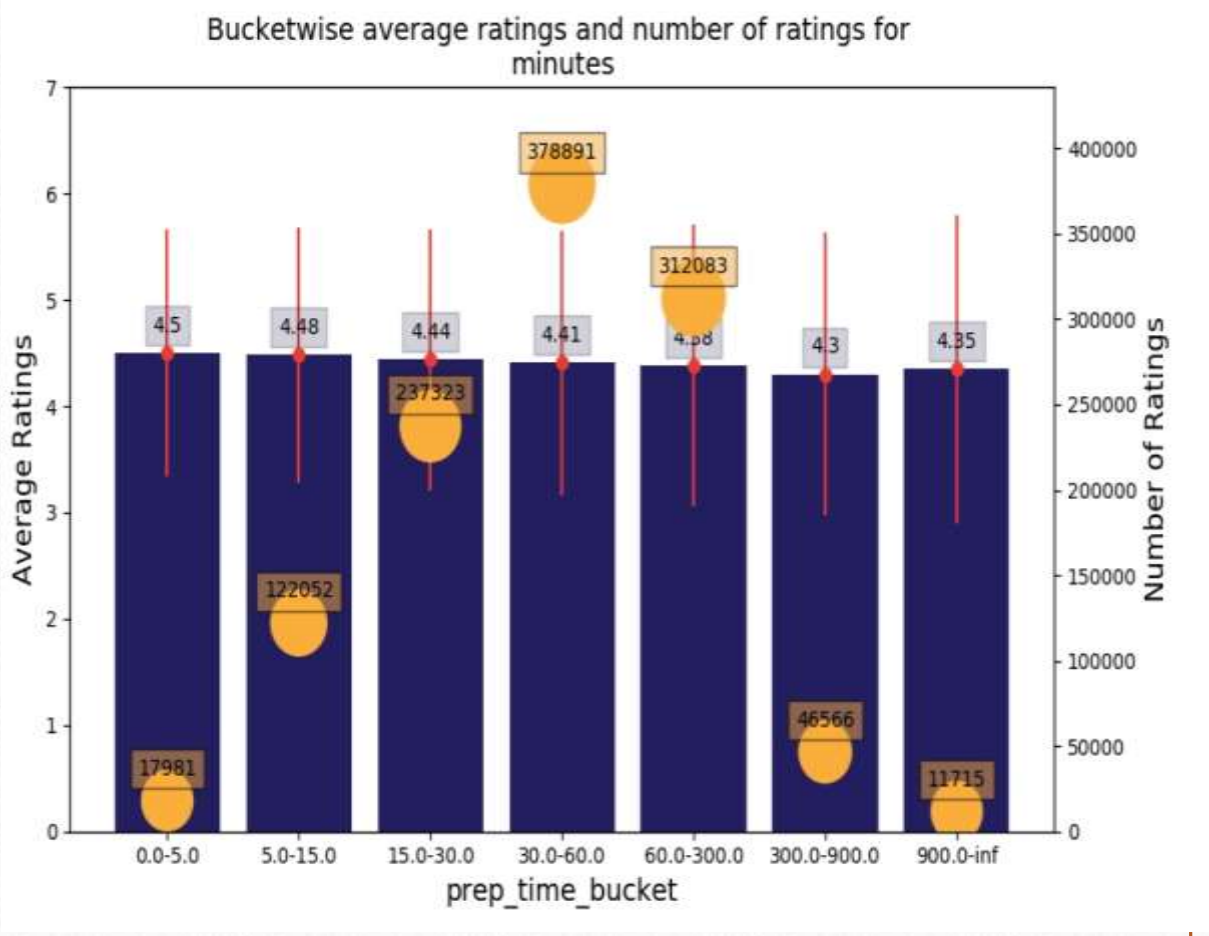
[Review Time Since Submission]

Recipes more than 6 years old are rated low



EDA

Preparation time
Somewhat relevant
Low prep time is preferred

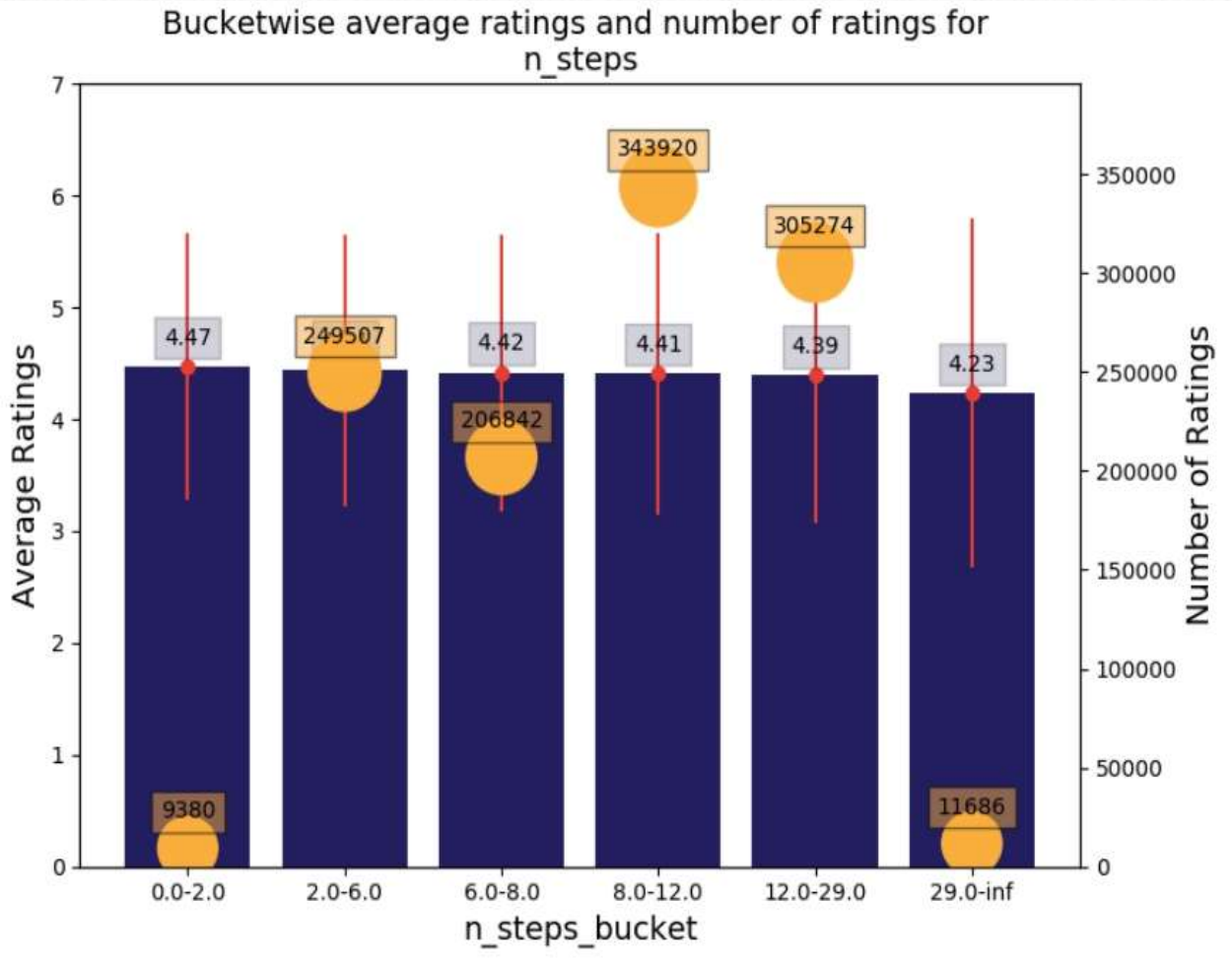


Number of steps

Clearly relevant

Recipes with less than 2 steps are rated high

Recipes with more than 29 steps are rated very low



Conclusion and Recommendations

- In this analysis of recipe data shows that the review time since submission and the number of steps, preparation time and number of ingredients are important factors in determining the rating of a recipe.
- Recipes that are reviewed by users after a long time from the submission date, have less number of steps, less preparation time and less number of ingredients tend to have high ratings of 5.
- In contrast, the number of ingredients in a recipe is not found to be relevant to the rating. Similarly, the nutrition columns such as calories, fat, sugar, sodium, protein, and fat. and per serving are not found to be relevant in determining the rating of a recipe.
- The findings of this analysis can be used to inform decisions about recipe development and presentation to users in order to meet their preferences

Thank You!

Meraj Alam
Neha Netkar