RECIPE RECOMMENDER ASSIGNMENT EDA

SUBMITTED BY:

KESHAV ATCHUTUNI

PRANAV KESHAV

MANU SMRITI

Problem Statement

Objective: Develop a recommender system to suggest recipes based on user preferences and the recipe they are currently viewing.

Potential Benefits:

- Increase user engagement.
- Create additional business opportunities.
- Directly impact the website's revenue.

Challenges:

- Building the system from scratch requires significant time and effort.

Task:

- Analyze available data.
- Generate relevant features for the recommender system.

TASK LIST

Task 1: Read the data

Task 2: Extract individual features from the nutrition column.

Task 3: Standardize the nutrition values.

Task 4: Convert the tags column from a string to an array of strings.

Task 5: Read the second data file

Task 6: Create time-based features.

Feature Extraction:

1. The Assignment is focused on extracting features from the data to build a recipe recommender system.
2. As we have observed, when the Spark compiler reads the nutrition column from the raw_recipes_df DataFrame, it is treated as a string column instead of an array of float values. However, each row in the nutrition column contains seven values representing different nutrition information. Our task is to extract these individual values and create seven separate columns named calories, total fat (PDV), sugar (PDV), sodium (PDV), protein (PDV), saturated fat (PDV), and carbohydrates (PDV).
3. Make nutrition-per-100 calorie columns
4. Creating nutrition-per-100 calorie columns: Like putting nutrition values on a level playing field, we standardize them to relative terms by considering 100 calories as the reference point.
5. Naming convention: Original column name "total fat (PDV)", column name after conversion "total_fat_per_100_cal"

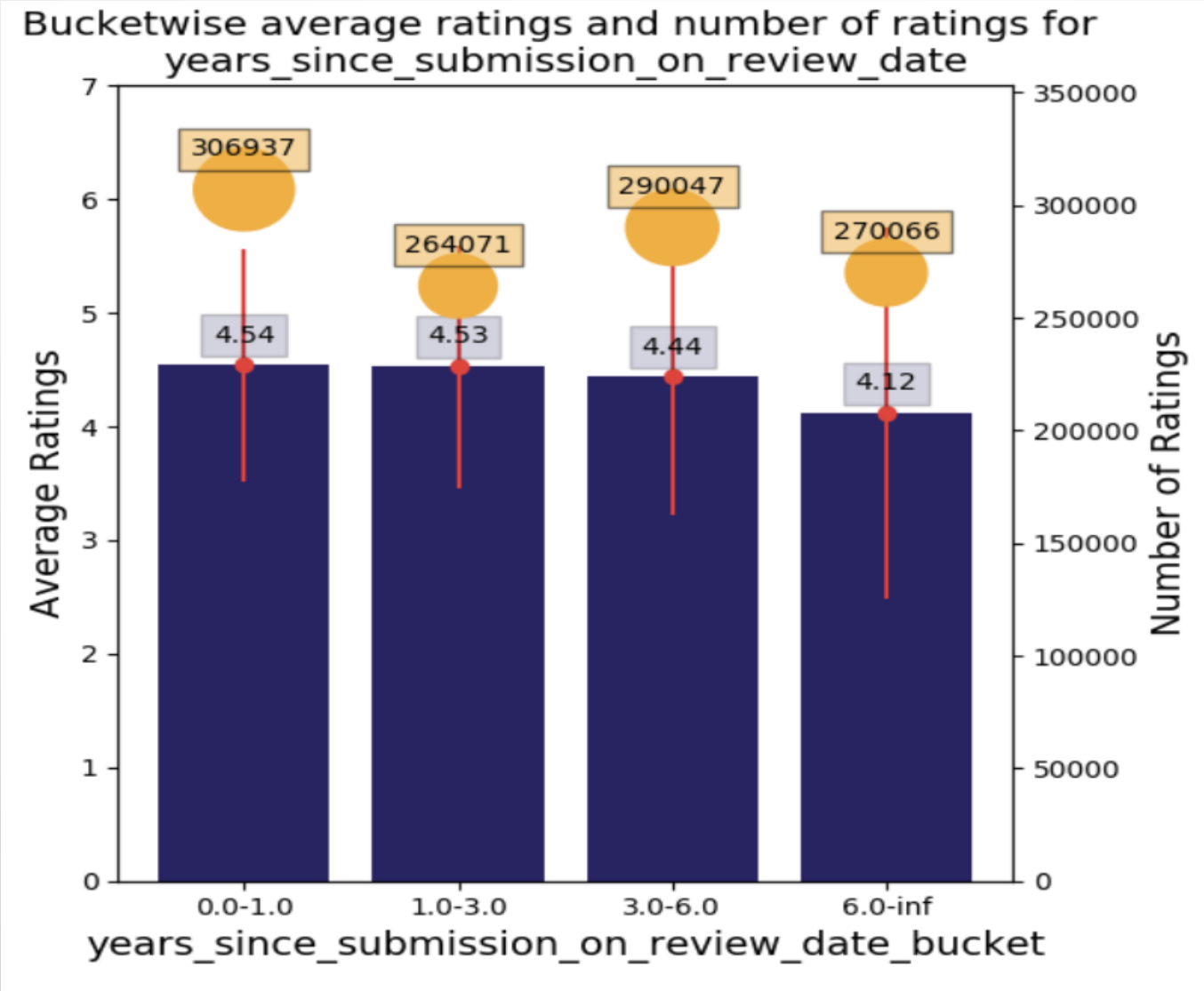
Exploratory Data Analysis:

The first approach to EDA is to cover all possibilities for all columns.

Bucketing & Cleaning Numerical Features

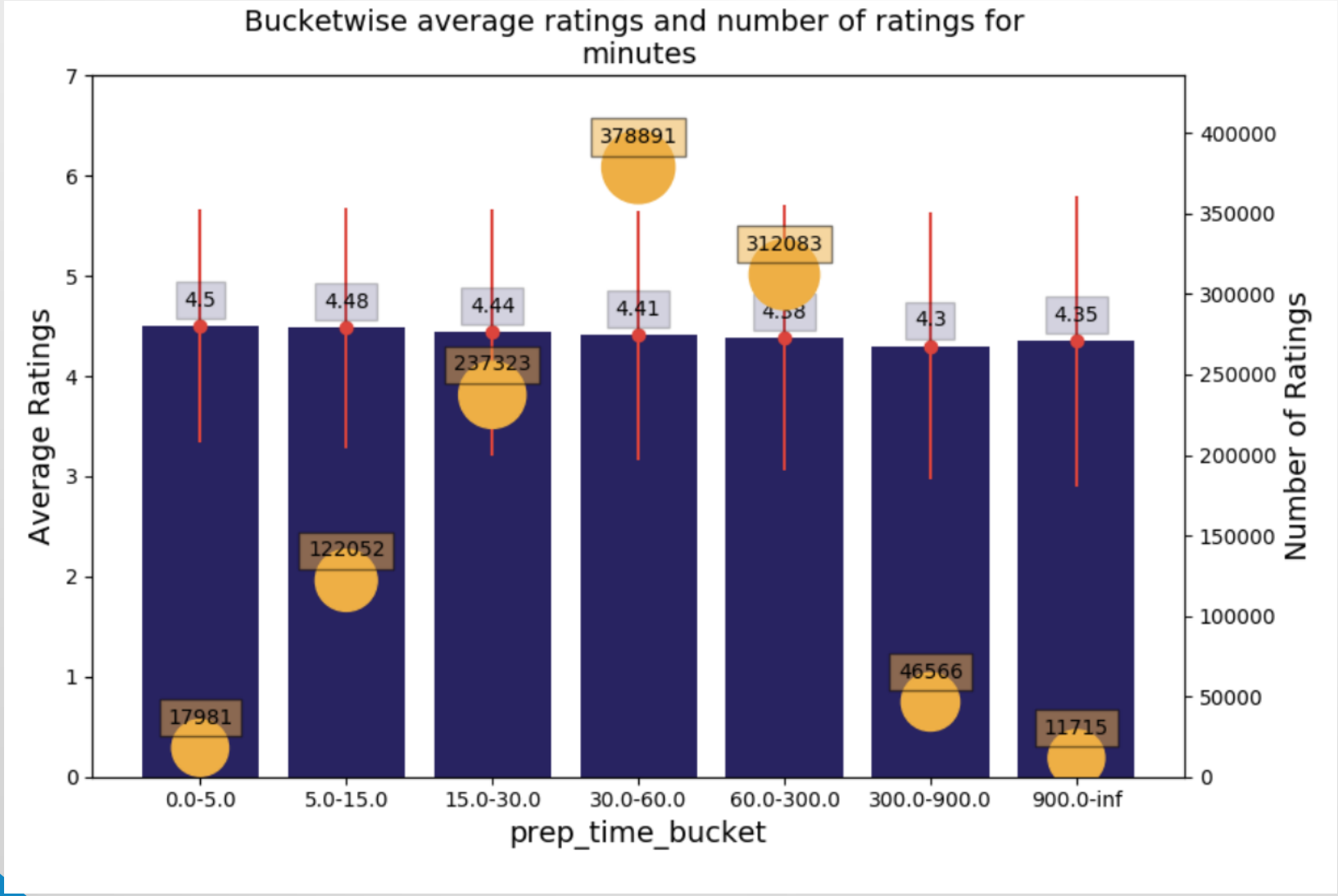
1. Years Since Submission On Review Date

Recipes more than 6 years old are rated low.



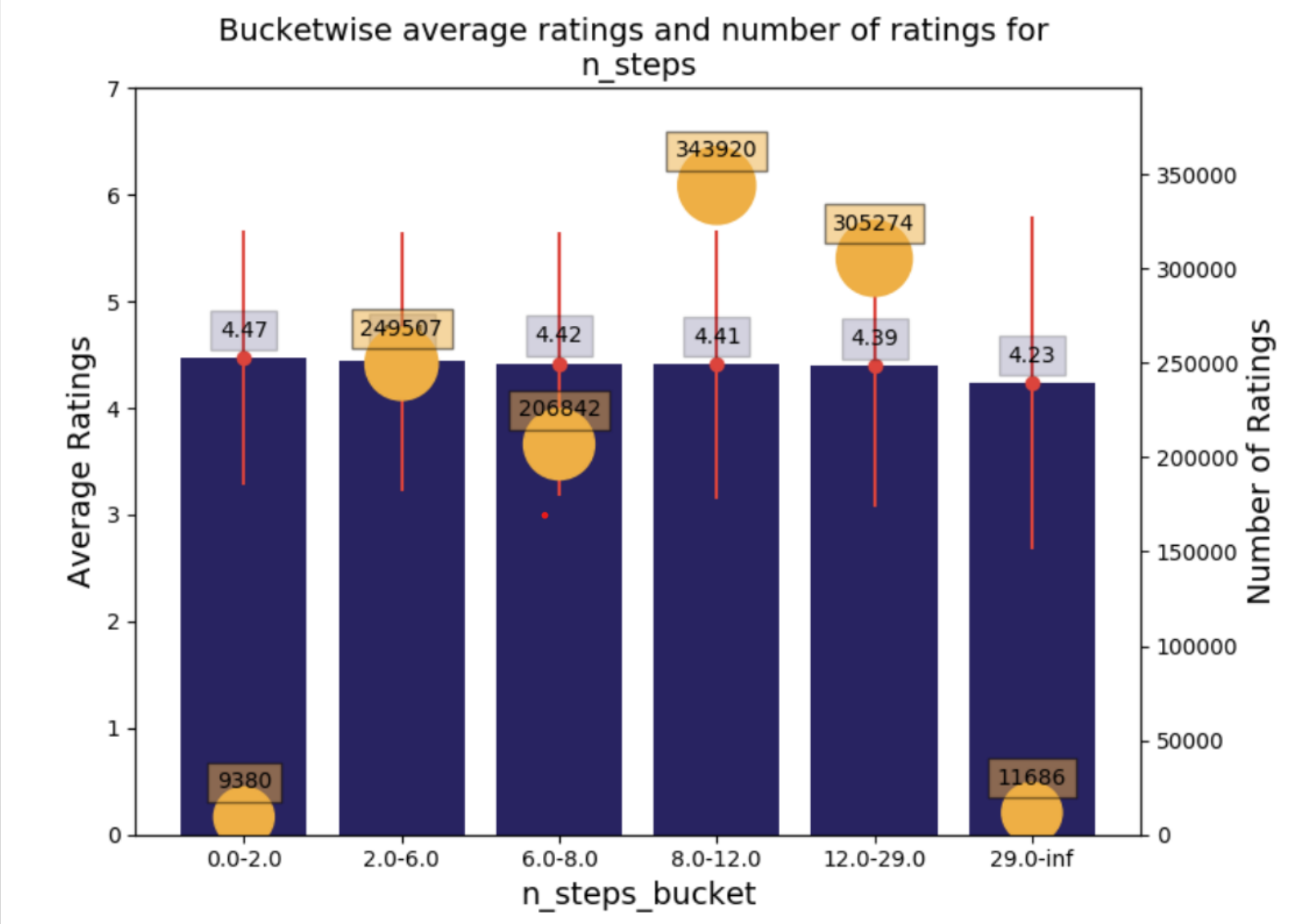
2. **Minutes**

A shorter preparation time is favored.



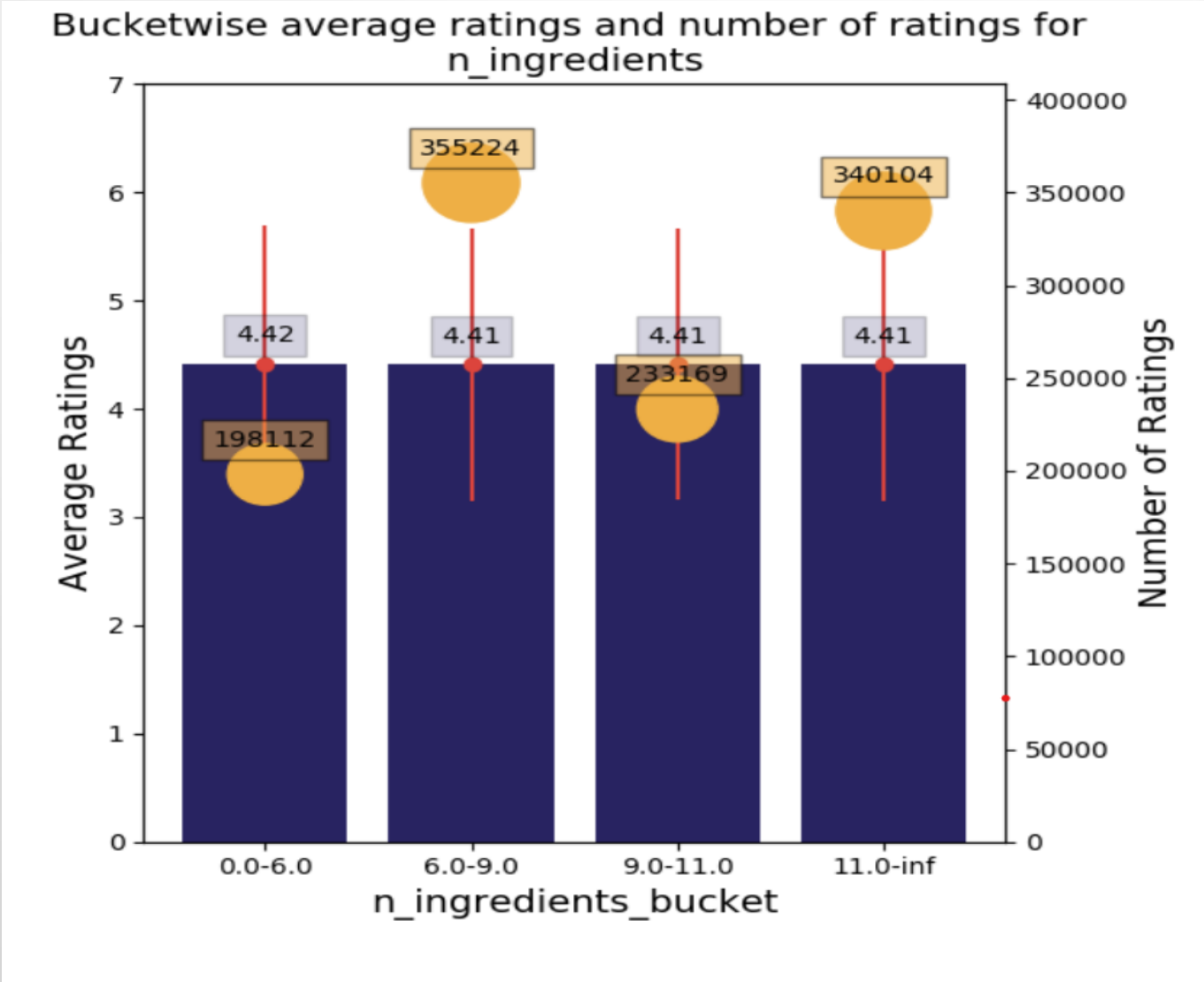
3. N-steps

a) Recipes with fewer than 2 steps receive high ratings.
b) Recipes with more than 29 steps receive very low ratings.



4. N_ingredients

- Insignificant



5. Nutrition Coloumns

- ✓ calories - Calories per serving seems irrelevant
- ✓ fat (per 100 cal) - Calories per serving seems irrelevant
- ✓ sugar (per 100 cal) - Calories per serving seems irrelevant
- ✓ sodium (per 100 cal) - Calories per serving seems irrelevant
- ✓ protein (per 100 cal) - Calories per serving seems irrelevant
- ✓ sat. fat (per 100 cal) - Calories per serving seems irrelevant
- ✓ carbs (per 100 cal) - Calories per serving seems irrelevant

Adding user level average features

More Features:

high_ratings = 5 rating

- user_avg_years_betwn_review_and_submission_high_ratings
- user_avg_prep_time_recipes_reviewed_high_ratings
- user_avg_n_steps_recipes_reviewed_high_ratings
- user_avg_n_ingredients_recipes_reviewed_high_ratings

CONCLUSION

1. In these notebooks, we have conducted exploratory data analysis (EDA) and feature extraction for the recipe recommendation system.
2. Developing an efficient recommender system is crucial for food.com to enhance user engagement and boost revenue.
3. Tailoring recipe recommendations based on user preferences and the current recipe will help maintain user interest on the site.
4. Increased user engagement creates potential for partnerships and promotions, leading to higher revenue.
5. Analyzing data and generating meaningful features are essential steps in building a successful recommendation engine.
6. Techniques such as collaborative filtering and content-based filtering will be used to provide precise recommendations.
7. Performance evaluation using metrics like precision, recall, and accuracy will help refine the system.
8. A dependable recommender system benefits users and supports the success of food.com.
9. As an ML engineer, you have the chance to influence user engagement and revenue generation through this system.