Recipe Recommender Assignment EDA

By

Ashish Singh Pawar

Bhavani Ramachandran

Abhishek Rajesh Dambe

Recipe Recommender Assignment
EDA

Problem Statement:

Design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at for food.com by Extracting features from data. The recommendation engine is a way to increase the website's user engagement. If a user is shown relevant recipes, they are more likely to spend more time on site reading about recipes. Higher user engagement will likely result in morebusiness opportunities like collaborations, promotions, etc. The performance of a recommendation engine will significantly impact the revenue your recipe site can generate.

BUSINESS OBJECTIVE:

- **This recipe involves analyzing user interactions with recipes to better grasp user preferences and discern patterns that could enhance recipe recommendations. This entails examining factors like review time since submission, preparation duration, number of steps, and ingredients to determine their correlation with high ratings. These findings can then guide the development of a more effective recipe recommendation algorithm, ultimately enhancing user engagement and satisfaction with the recipe platform or app. Moreover, such analysis can uncover areas for potential enhancement in recipe content and presentation, such as simplifying steps or ingredient lists. Ultimately, the objective is to enhance user experience, thereby bolstering customer retention and acquisition.**

- from pyspark.sql import SparkSession
- from pyspark.sql import SparkSession
- spark = SparkSession.builder.appName("Basics").getOrCreate()
- from pyspark.sql import functions as F
- Import for typecasting columnsfrom pyspark.sql.types
- import IntegerType,BooleanType,DateType,FloatType,StringType
- from pyspark.sql.types import ArrayType

◦ This code imports various functions and types from the PySpark library, which is used for working with data in the Apache Spark framework.

◦ we can use these types and functions in our code to manipulate and analyze data stored in Spark DataFrames

◦ We have included some test cases given below. We have complete the tasks

Solution Methodology
Data cleaning and data manipulations

- Comprehensive Solution to All Tasks
- Data Reading
- Compilation of Nutrition Columns
- Extraction of Distinct Features from the Nutrition Column
- Application of String Operations to Eliminate Square Brackets from the Nutrition Column
- Segregation of the Nutrition Column into Seven Distinct Columns and Conversion of the New Columns into Floating-Point Values
- Segmented Nutrition Column
- Below are Provided Test Cases for Verification of Task Completion Accuracy

DATA MANIPULATION

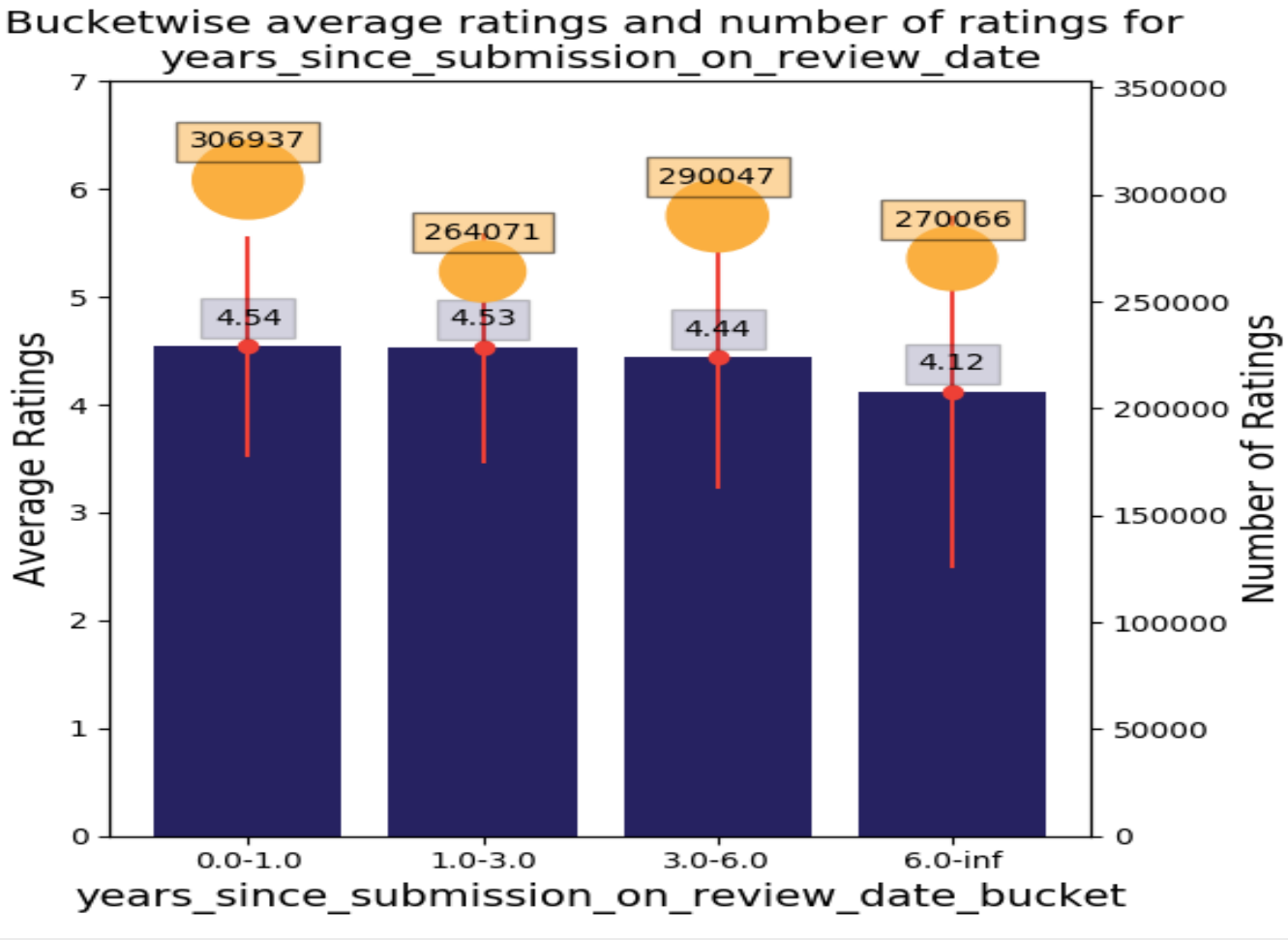
- Normalize the Nutrition Values
- By converting the nutrition values from absolute to relative terms, we ensure that portion size does not influence the analysis.
- Standardize all nutrition columns to per 100 calories. Test cases are provided below for verification of task completion accuracy.
- Complete the code in the following cell.
- Transform the tags column from a string to an array of strings.
- Merge Recipe Data with Review Data and Load the second data file.
- Generate time-based features.
- Save the data we've generated so far in a parquet file.

('s3://recipecasestudy/interaction_level_df_processed.parquet')

- **Has a header row and that the data types for all columns should be inferred automatically. The file is located at "s3a://raw-recipes-clean**

- 1.('s3://receipecasestud/interaction_level_df_processed.parquet')
- 2.('s3a://upgradfoodrecsysdir/interaction_level_df_postEDA.parquet')
- 3.('s3a://upgradfoodrecsysdir/interaction_level_df_ModelReady.parquet')
- **Such values and Null values are treated as Not Declared and used for further analysis • Numerical Missing values have been dropped • Outlier Treatment of TotalVisits and Page Views Per Visit. • Observed that major part of null values in "Page Views Per Visit", "TotalVisits" are Converted. So, imputing • median values of them to null values.**

EDA

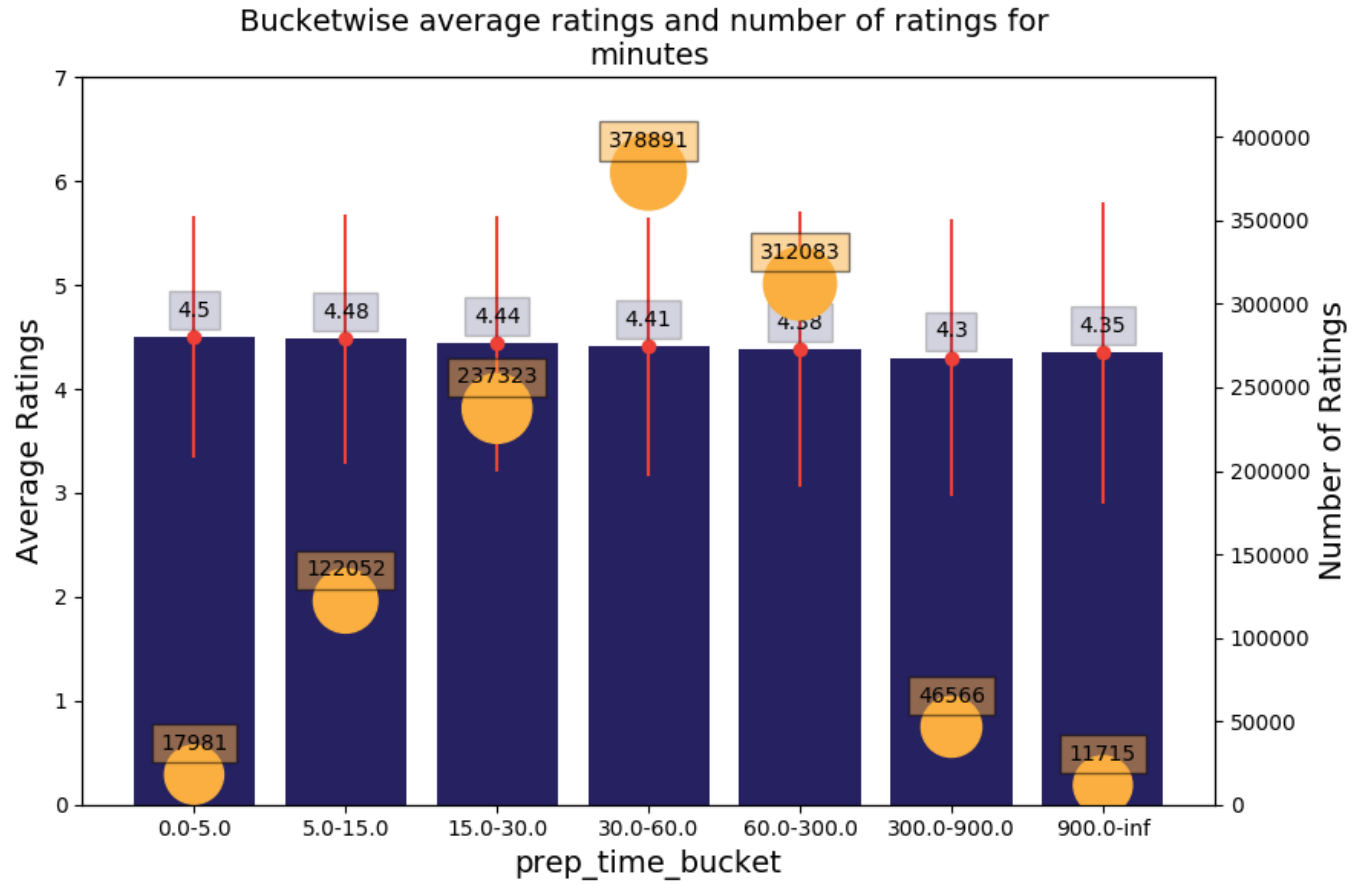


Years since submissionon review date

[Review Time Since Submission]

Receipe more than 6 years old are rated low

EDA



[prep time]

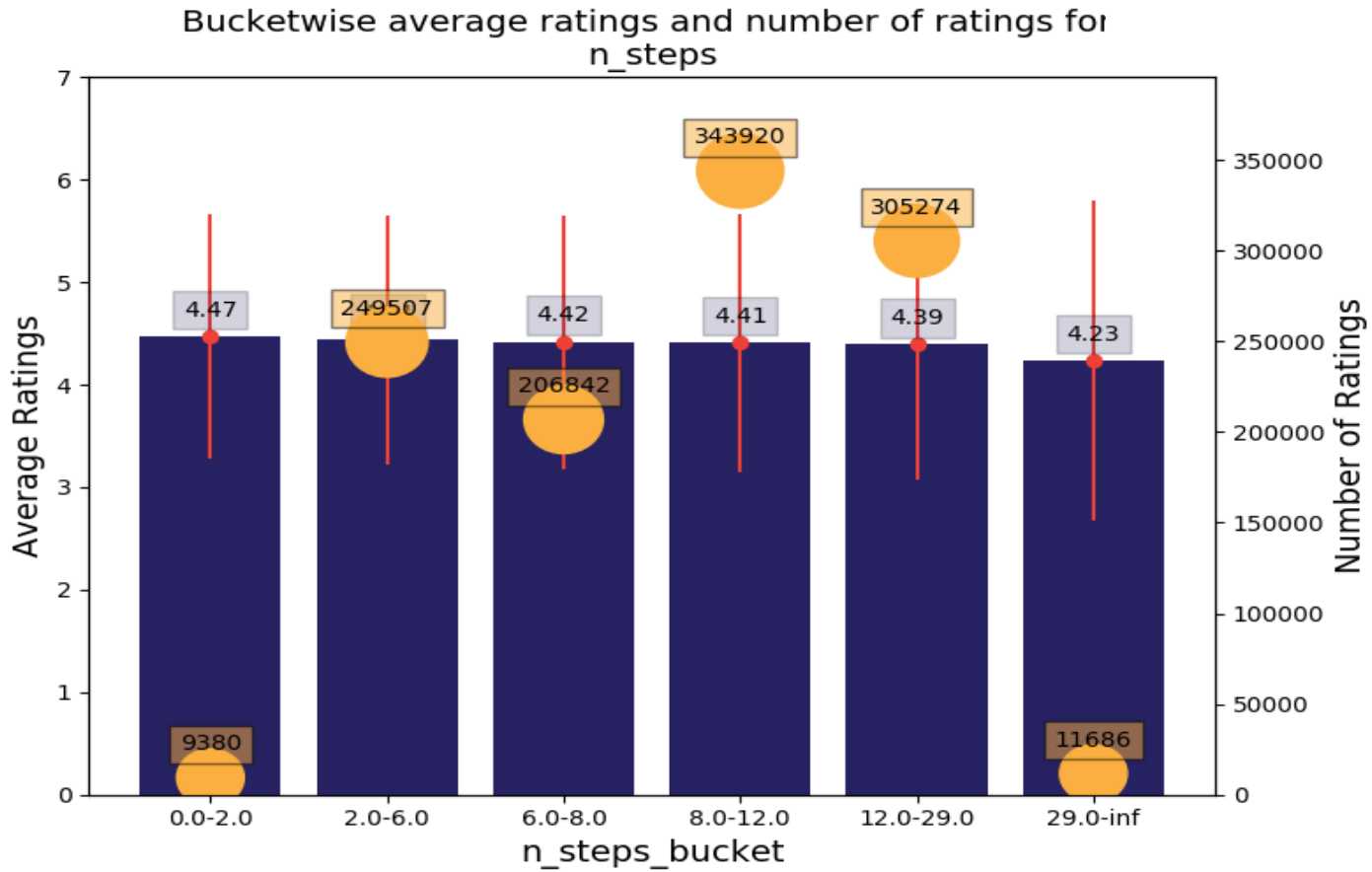
- Somewhat relevant

- Low prep time is preferred

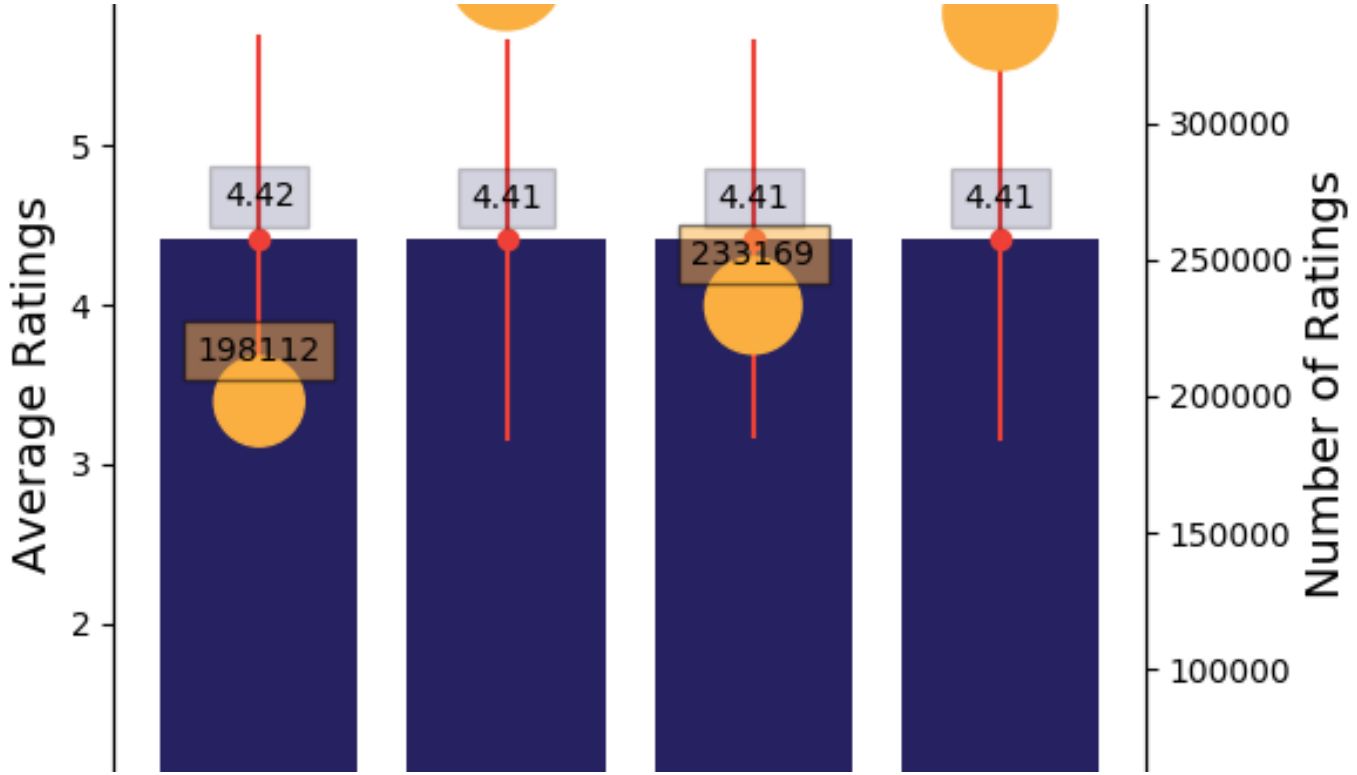
EDA

n_steps:

- **Clearly relevant**
- **Recipes with less than 2 steps are rated high**
- **Recipes with more than 29 steps are rated very low**



EDA



Number of ingredients

Not Relevant

EDA

- **Nutrition Column:**
  •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

EDA

- **More Features:**
- **High ratings = 5 rating**
- **User average years between review and submission high ratings**
- **User average Preparation time recipes reviewed high ratings**
- **User average number of steps recipes reviewed high ratings**
- **User average number of ingredients recipes reviewed high ratings**

EDA

Top numbers most rated tags

individual_tag
preparation
time-to-make
course
dietary
main-ingredient
easy
occasion
equipment
cuisine
low-in-something
main-dish
60-minutes-or-less
number-of-servings

EDA

Bottom and least rated tags:

individual_tag
lamb-sheep-main-dish
black-bean-soup
Throw the ultimat...
chicken-stews
desserts-easy

EDA

Top and rated tags:

individual_tag
breakfast-potatoes
pork-loin
middle-eastern-ma...
Throw the ultimat...
desserts-easy

Conclusion and Recommendations:

The analysis of recipe data indicates that certain factors significantly influence the rating of a recipe. Specifically, factors such as review time since submission, number of steps, preparation time, and number of ingredients play pivotal roles in determining a recipe's rating. Recipes that receive reviews long after their submission date, have fewer steps, shorter preparation times, and fewer ingredients tend to garner higher ratings, typically achieving a rating of 5.

Interestingly, the number of ingredients alone does not appear to strongly correlate with the recipe's rating. Additionally, nutrition-related factors such as calories, fat content, sugar content, sodium levels, protein content, and fat per serving do not seem to significantly affect the recipe's rating.

These insights can be valuable for guiding decisions related to recipe development and presentation to users. By aligning with user preferences as revealed by these findings, recipe platforms can enhance user satisfaction and engagement. This could involve optimizing recipe content and presentation to emphasize factors such as shorter preparation times and fewer steps, while also ensuring the nutritional aspects meet basic standards without necessarily impacting the rating.

THANK YOU