Assignment : Recommender system to recommend recipes

Group Members' name :

□ Rachelin Sujae P
□ Ramyaseetha V
□ Vishal Kumar

Objective:

The primary goal is to create a system that recommends recipes to users based on their preferences and the recipes they are currently viewing by performing Exploratory Data Analysis and Feature Extraction from the raw data (www.food.com).

Problem Statement:

The significance of this recommendation engine should not be underestimated. It plays a crucial role in enhancing user engagement on the www.Food.com website. When users are presented with recipes that align with their interests, they are more likely to spend additional time exploring the platform, reading about recipes, and interacting with the content. This increased user engagement opens up various business opportunities, such as collaborations and promotions, which can ultimately lead to a boost in revenue for the recipe site.

Data used in assignment

- Raw_recipes.csv File Overview: The Raw_recipes.csv file contains comprehensive recipe-related information, with each row representing a distinct recipe. The data fields within this file encompass various recipe attributes and details.

- RAW_interactions.csv File Overview: The RAW_interactions.csv file is dedicated to user interactions with recipes. Each row corresponds to a user's review of a specific recipe. Notably, users may review multiple recipes, and recipes can receive reviews from numerous users. The uniqueness of each row is ensured by the combination of user_id and reviewer_id. Additional attributes related to these interactions are also included in RAW_interactions.csv.

FEATURE EXTRACTION AND EDA

The below codes to import various libraries and functions which is used for working with data in Apache Spark framework.

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

TASKS PERFORMED

Task 1: a) Read RAW_recipes.csv from S3 bucket. b) Ensured each field has the correct data type.
Task 2: Extracted individual features from the nutrition column.
Task 3: Standardized 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.
Task 7: Processing Numerical Columns and EDA
Task 8: Create user-level features
Task 9: Create tag-level features

Treating the nutrition column.

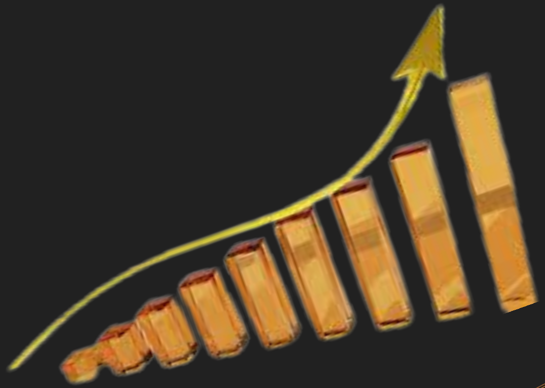
Extracting Individual Features

- The nutrition column is currently in string format.
- It should be converted into an array of float values.
- Each row in the nutrition column contains seven values representing nutrition information.
- These values need to be separated into seven individual columns: calories, total fat PDV, sugar PDV, sodium PDV, protein PDV, saturated fat PDV, and carbohydrates PDV.

Standardizing Nutrition Values

- Nutritional values in absolute terms can vary based on serving size.
- The goal is to standardize the nutrition values by converting them to per 100 calories.

To address the issue of nutritional variation due to serving size discrepancies, it's important to standardize the nutrition values. This involves converting the nutritional values to a consistent measurement per 100 calories, ensuring a fair comparison across different recipes and serving sizes.



nutrition columns

- 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

Converting a string to an array of strings

Problem:

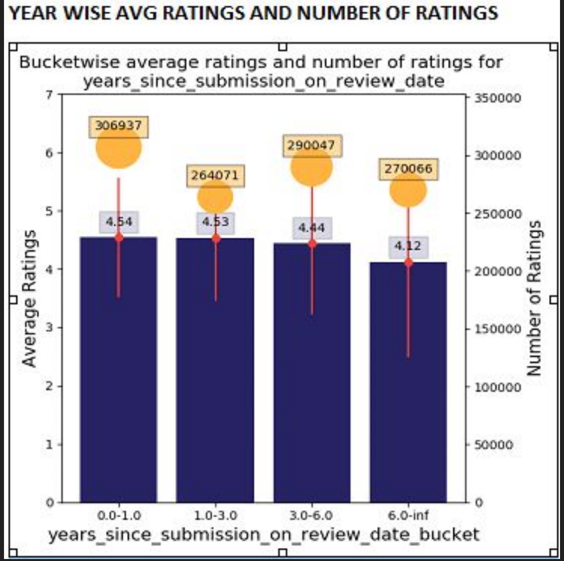
The `tags` column in a dataset is a string, but it contains an array of strings. This can be problematic because it makes it difficult to work with the data. For example, it can be difficult to filter or sort the data by tag.

Solution:

We can use the `split()` method to convert the `tags` column from a string to an array of strings. The `split()` method takes a delimiter as an argument, and it splits the string into a list of strings at each occurrence of the delimiter.

Converting a string to an array of strings is a simple but effective way to improve the quality of your data. It makes the data easier to work with, more expressive, and more efficient.

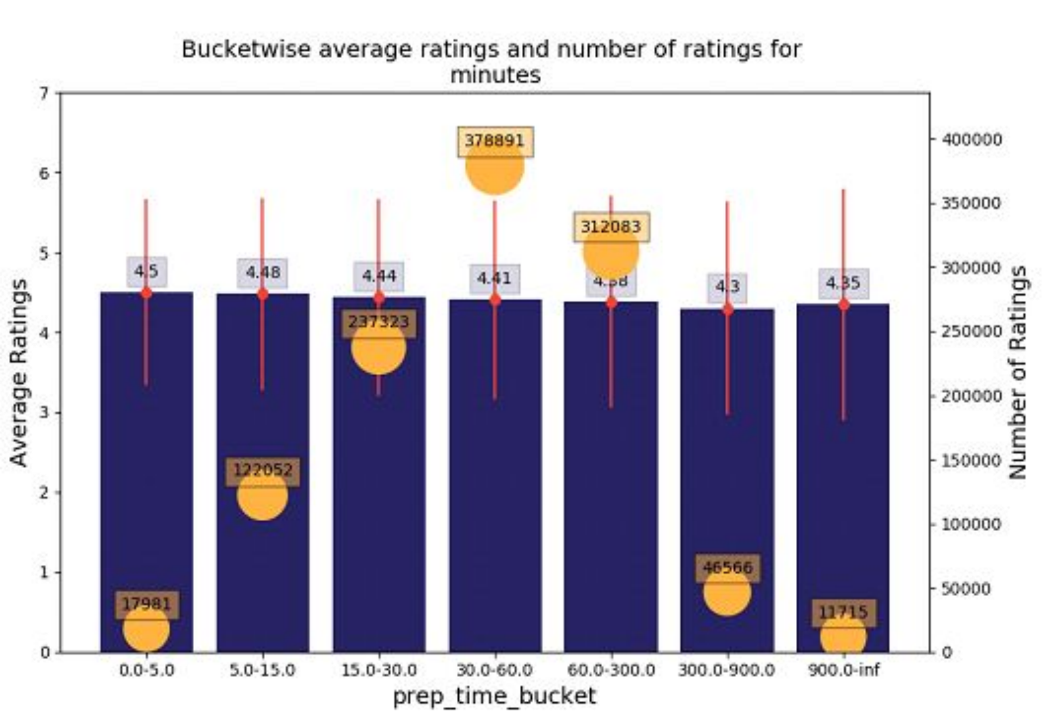
YEAR WISE AVG RATINGS AND NUMBER OF RATINGS



Inference:

- The preparation time exhibits a certain degree of correlation among themselves.
- Recipes with shorter preparation times tend to receive the highest ratings.

PREPARATION TIME-BASED AVG RATINGS AND NUMBER OF RATINGS



Inference:

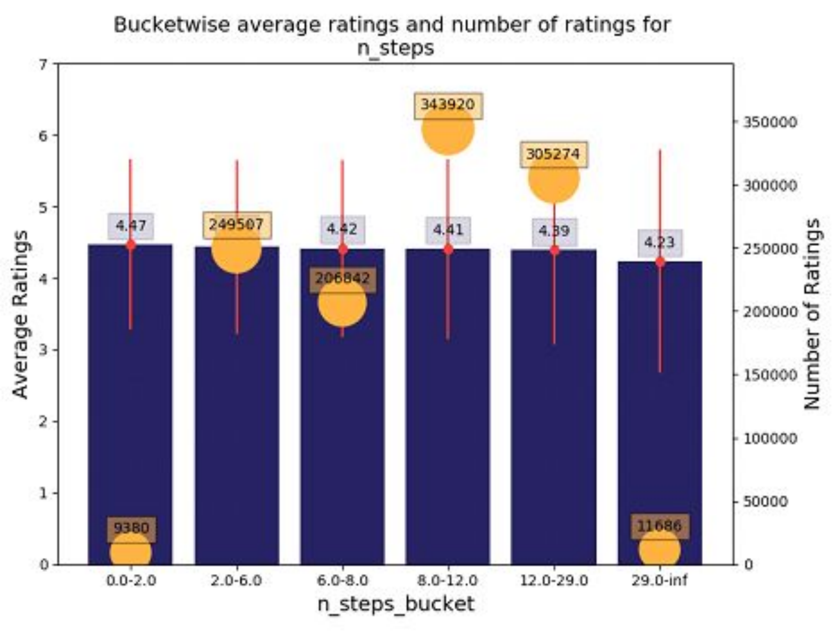
- Recipes that have been in existence for over six years tend to receive lower ratings.

COUNT OF RECIPE STEPS Vs AVG RATINGS AND NUMBER OF RATINGS

Inference:

•Recipes containing fewer than 2 steps tend to receive high ratings.

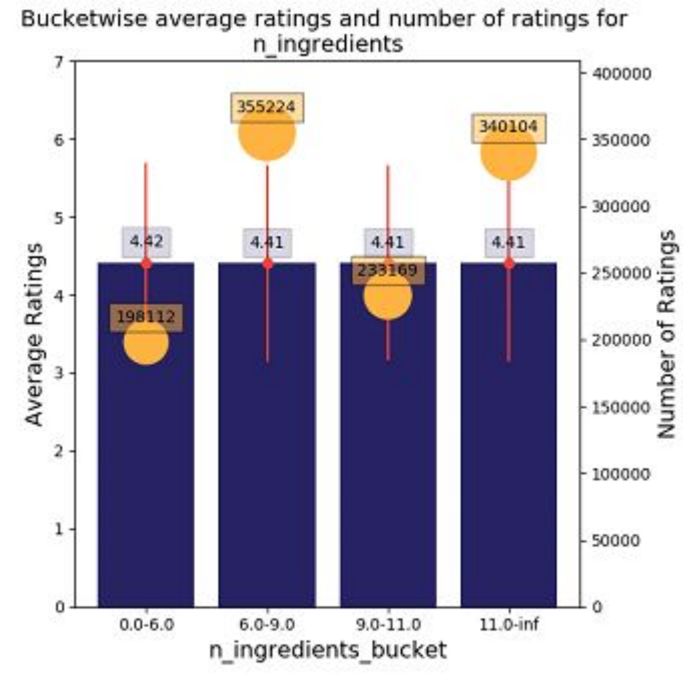
•Conversely, recipes involving more than 29 steps tend to receive very low ratings.



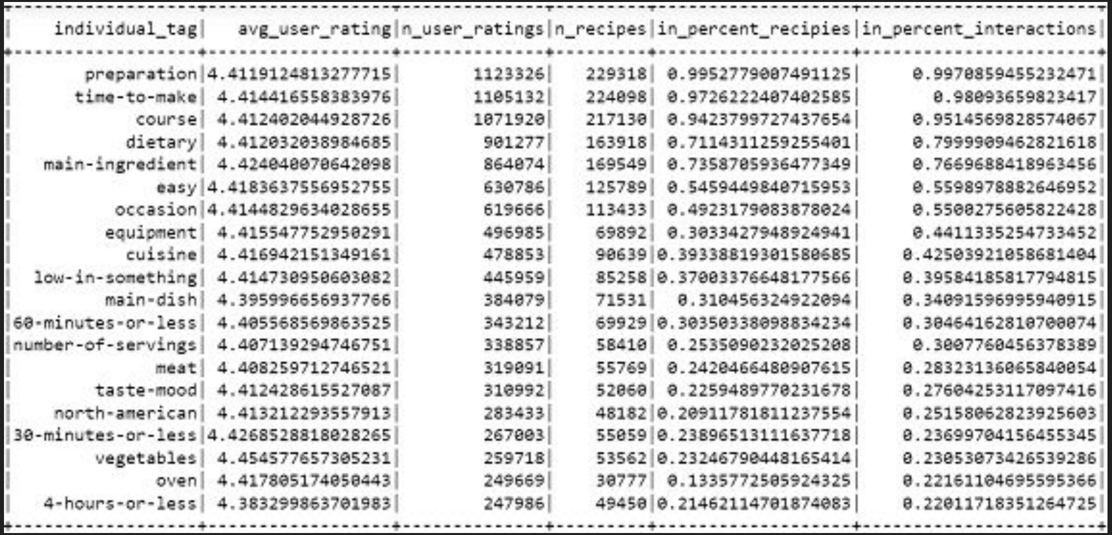
INGREDIENT COUNT Vs AVG RATINGS AND NUMBER OF RATINGS

Inference:

•There is no correlation between the ingredients and ratings

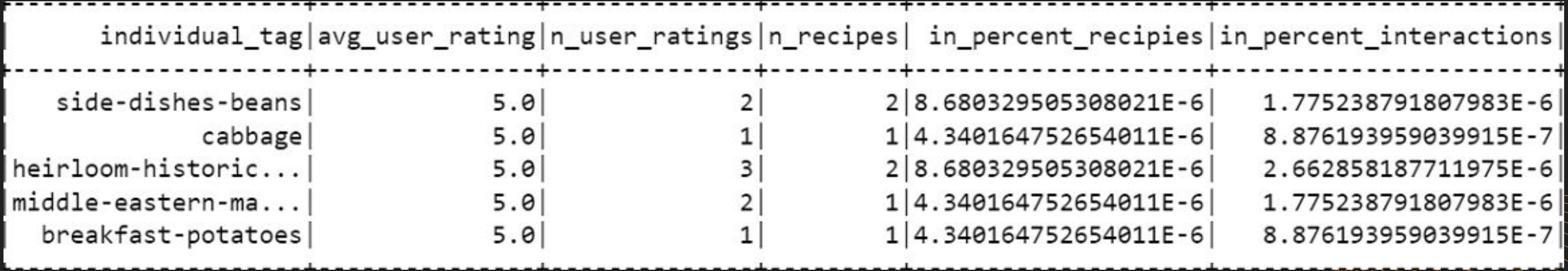


Tags having top 20 highest count of user rating



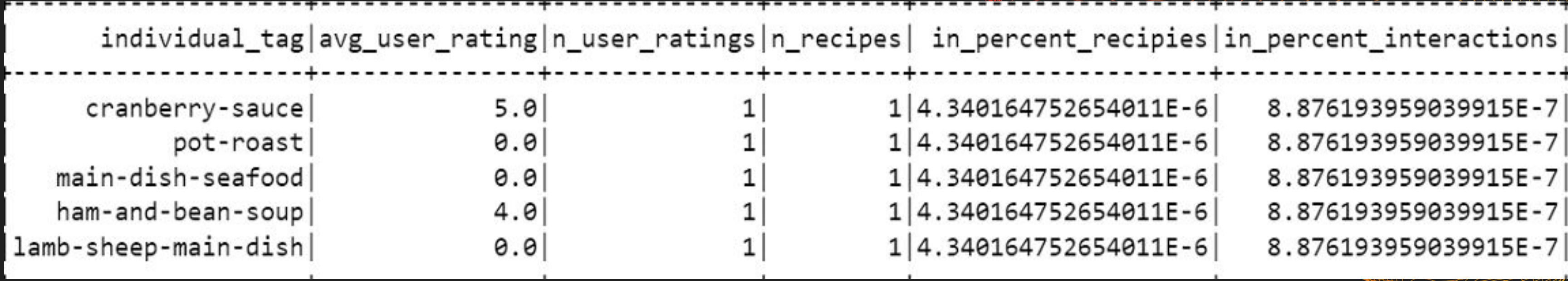
```
+------------------+------------------+--------------+---------+-------------------+-----------------------+
|    individual_tag|   avg_user_rating|n_user_ratings|n_recipes|in_percent_recipies|in_percent_interactions|
+------------------+------------------+--------------+---------+-------------------+-----------------------+
|       preparation|4.4119124813277715|       1123326|   229318| 0.9952779007491125|     0.9970859455232471|
|      time-to-make| 4.414416558383976|       1105132|   224098| 0.9726222407402585|       0.98093659823417|
|            course| 4.412402044928726|       1071920|   217130| 0.9423799727437654|     0.9514569828574067|
|           dietary| 4.412032038984685|        901277|   163918| 0.7114311259255401|     0.7999909462821618|
|   main-ingredient| 4.424040070642098|        864074|   169549| 0.7358705936477349|     0.7669688418963456|
|              easy|4.4183637556952755|        630786|   125789| 0.5459449840715953|     0.5598978882646952|
|          occasion|4.4144829634028655|        619666|   113433| 0.4923179083878024|     0.5500275605822428|
|         equipment| 4.415547752950291|        496985|    69892| 0.3033427948924941|     0.4411335254733452|
|           cuisine| 4.416942151349161|        478853|    90639|0.39338819301580685|    0.42503921058681404|
|  low-in-something| 4.414730950603082|        445959|    85258|0.37003376648177566|    0.39584185817794815|
|         main-dish| 4.395996656937766|        384079|    71531|  0.310456324922094|    0.34091596995940915|
|60-minutes-or-less| 4.405568569863525|        343212|    69929|0.30350338098834234|    0.30464162810700074|
|number-of-servings| 4.407139294746751|        338857|    58410| 0.2535090232025208|     0.3007760456378389|
|              meat| 4.408259712746521|        319091|    55769| 0.2420466480907615|    0.28323136065840054|
|        taste-mood| 4.412428615527087|        310992|    52060| 0.2259489770231678|    0.27604253117097416|
|    north-american| 4.413212293557913|        283433|    48182|0.20911781811237554|    0.25158062823925603|
|30-minutes-or-less|4.4268528818028265|        267003|    55059|0.23896513111637718|    0.23699704156455345|
|        vegetables| 4.454577657305231|        259718|    53562|0.23246790448165414|    0.23053073426539286|
|              oven| 4.417805174050443|        249669|    30777| 0.1335772505924325|    0.22161104695595366|
|   4-hours-or-less| 4.383299863701983|        247986|    49450|0.21462114701874083|    0.22011718351264725|
+------------------+------------------+--------------+---------+-------------------+-----------------------+
```

Top 5 rated tags



```
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
|      individual_tag|avg_user_rating|n_user_ratings|n_recipes| in_percent_recipies|in_percent_interactions|
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
|   side-dishes-beans|            5.0|             2|        2|8.680329505308021E-6|   1.775238791807983E-6|
|             cabbage|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|heirloom-historic...|            5.0|             3|        2|8.680329505308021E-6|   2.662858187711975E-6|
|middle-eastern-ma...|            5.0|             2|        1|4.340164752654011E-6|   1.775238791807983E-6|
|  breakfast-potatoes|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
```

Bottom 5 least rated tags



```
+--------------------+---------------+--------------+---------+--------------------+-----------------------
|      individual_tag|avg_user_rating|n_user_ratings|n_recipes| in_percent_recipies|in_percent_interactions|
+--------------------+---------------+--------------+---------+--------------------+-----------------------
|     cranberry-sauce|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7
|           pot-roast|            0.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7
|   main-dish-seafood|            0.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7
|   ham-and-bean-soup|            4.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7
|lamb-sheep-main-dish|            0.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7
```

Create Time-Based Features

Instructions for creating features that capture the time elapsed between a review and the date on which the recipe was submitted. This involves using the review_date and submitted columns after joining two data files.

Feature Creation

- Create features that calculate the time passed between a review and the recipe's submission date.
- Utilize the review_date and submitted columns for this calculation.

Advantages of Time-Based Features

- Time-based features provide insights into recipe popularity trends over time.
- They help identify the impact of seasonality on recipe reviews and engagement.
- These features enable tracking user behavior changes and adapting content accordingly.

Summary

- ❖ In this case study, a comprehensive feature extraction was conducted from two primary files, namely RAW_recipes.csv and RAW_interaction.csv. Our key achievements and steps undertaken include:
- ❖ Nutrition Data Enhancement: The nutrition array is successfully separated into seven distinct columns, aligning with the specified task requirements.
- ❖ Normalization of Nutritional Values: Recognizing the significant variation in absolute nutritional terms, we normalized the nutritional values on a per 100-calorie basis, ensuring a consistent and meaningful comparison.
- ❖ Tags Column Transformation: The tags column was transformed from a string format into an array of strings, aligning with the defined task objectives.
- ❖ Time Elapsed Feature Generation: New features were introduced to calculate the time elapsed between each review and the submission date of the corresponding recipe. This was achieved through merging and utilizing data from the review date and submitted columns.

Our analysis has revealed several crucial factors that significantly influence the rating of a recipe:

- Review Time Since Submission: The duration between a recipe's submission date and when it is reviewed plays a pivotal role. Recipes reviewed long after submission, with fewer steps, shorter preparation times, and fewer ingredients, tend to receive high ratings, often achieving a top rating of 5.

- Number of Ingredients: Surprisingly, we found that the number of ingredients in a recipe does not seem to have a substantial impact on its rating.

- Nutritional Data: Additionally, various nutritional attributes such as calories, fat, sugar, sodium, protein, and fat per serving were found to be unrelated to a recipe's rating.

Recommendations:

- Continue to monitor and maintain data quality and consistency, particularly when dealing with nutritional information and time-related features.
- Explore further data analysis and visualization techniques to gain deeper insights into the relationships between these extracted features and recipe ratings.
- Consider employing machine learning models to predict and understand how these extracted features influence recipe ratings, potentially leading to improved recommendations and user engagement.
- After conducting feature extraction and analysis, the resulting data has been securely stored in the S3 bucket (s3://demobucket13092023/RECIPE_CASESTUDY_DATA/). This data is now poised for utilization in the recipe recommendation system. It is expected that these insights will enhance the system's ability to recommend recipes that align with user preferences and ultimately lead to higher user satisfaction and engagement.