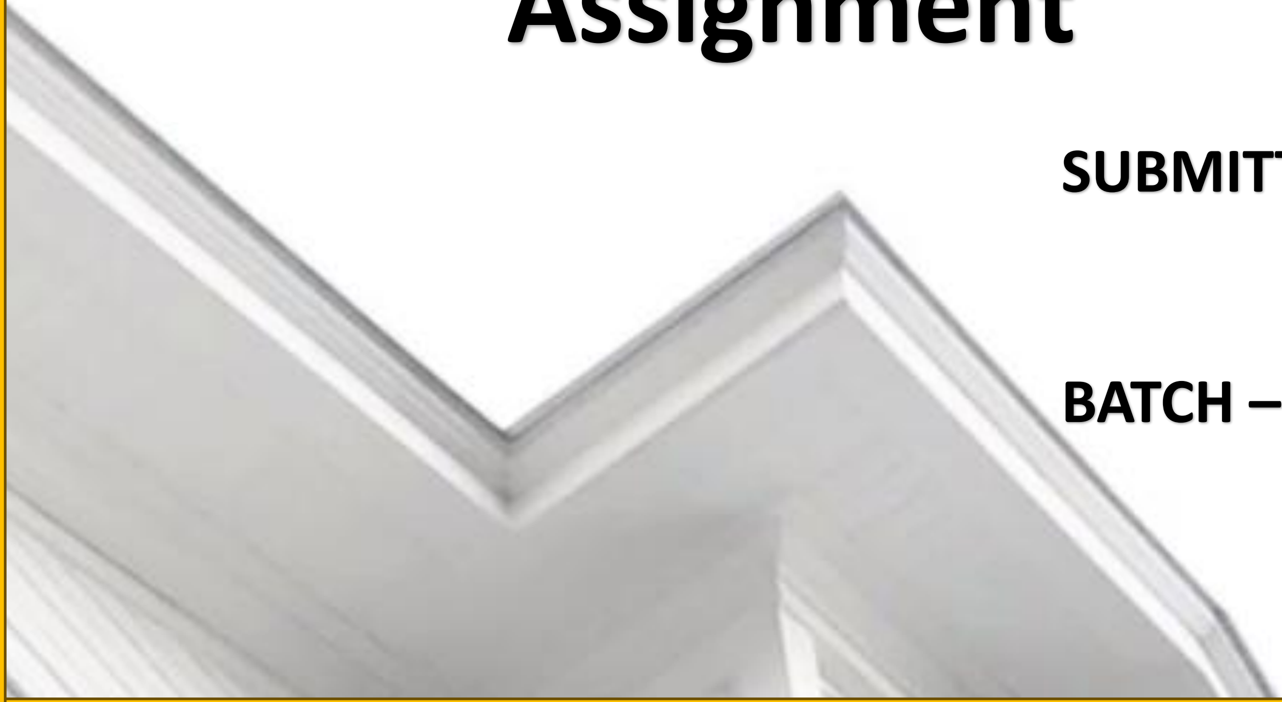
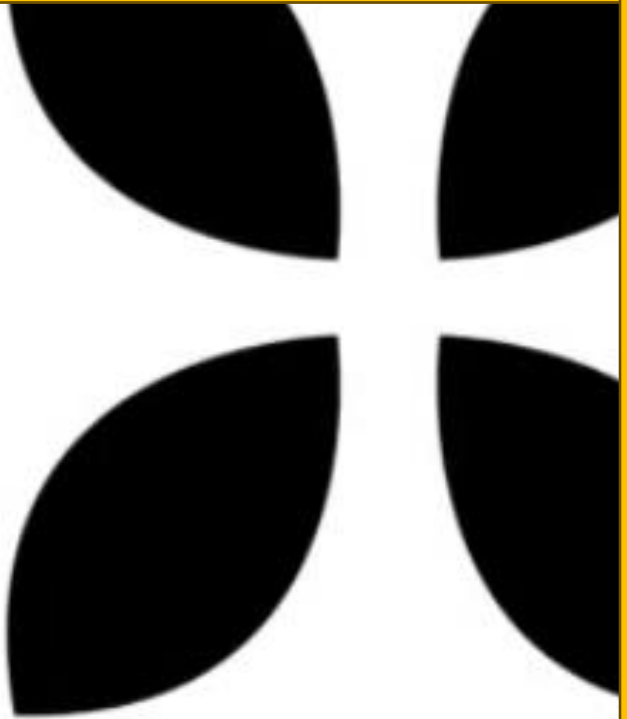
Recipe Recommender Assignment

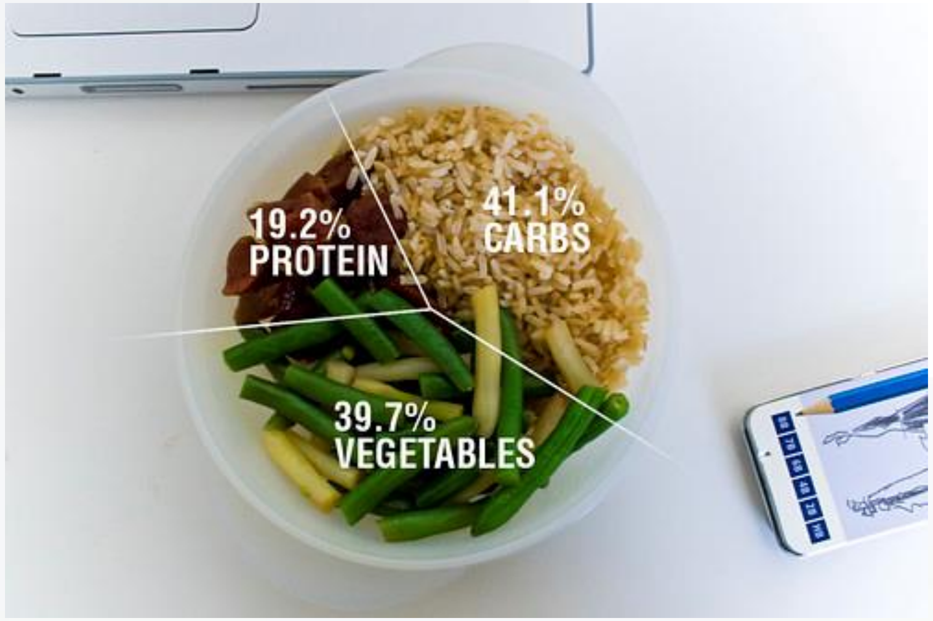




SUBMITTED BY –Ankita Sethi
Bharath Konda
Besty Boves

BATCH –DSC-56

Problem Statement



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.

BUSINESS OBJECTIVE

- The objective of this entire assignment is to perform Exploratory Data Analysis and feature extraction from the raw data.
- To identify user preferences and identify patterns that can be used to improve the recipe recommendations for users.
- This can be done by analyzing factors such as what all ingredients are required, number of steps, time of preparation, review time since submission and determining what overall factors are strongly related with high ratings.
- With this we can build recommendation algorithm which will automatically help us to increase user engagement and satisfaction.
- Our main goal is to improve user experience and increase customer retention.

Important Libraries

Some libraries which are imported to perform various tasks –

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Basics").getOrCreate()
from pyspark.sql import functions as F

# Import for typecasting columns
from pyspark.sql.types import
IntegerType,BooleanType,DateType,FloatType,StringType
from pyspark.sql.types import ArrayType

from pyspark.sql.functions import split,col
from pyspark.sql.types import ArrayType, StringType
```

TASKS PERFORMED

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.
Task 7: Processing Numerical Columns
Task 8: Create user-level features
Task 9: Create tag-level features

SOLUTION

- Reading the data.
- List of nutrition columns
- Extracting individual features from the nutrition column.
- Using string operations to remove the brackets from the nutrition column
- Splitting the nutrition column into seven columns and casting new columns to float values.
- Nutrition column split into multiple.
- Some test cases were used which can be used to check if you have completed the task correctly.

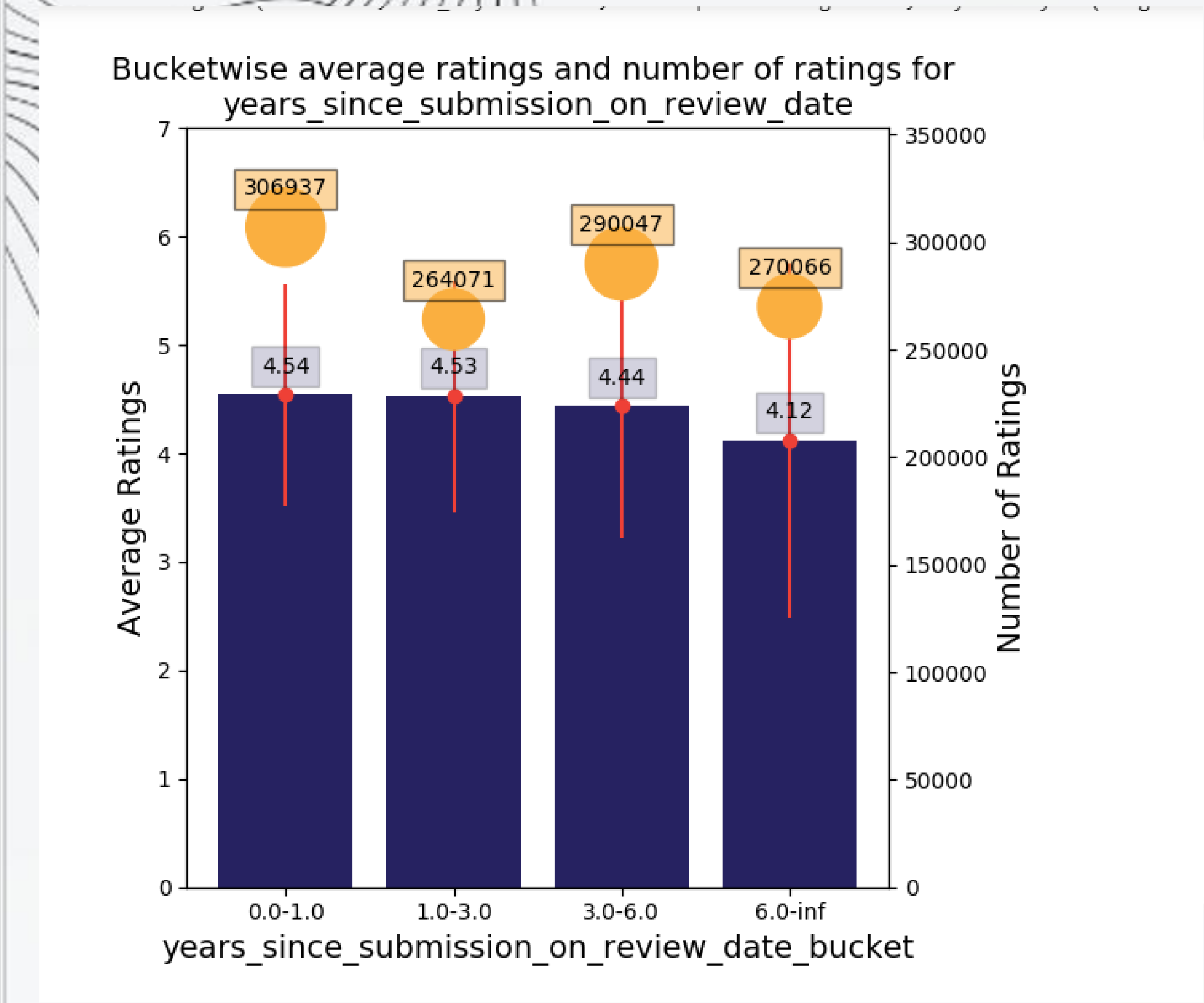
Standardizing the nutrition values

- By converting the nutrition values from absolute to relative terms, we are ensuring that portion size is not a factor in the analysis.
- All nutrition columns are standardized to per 100 calories.
- Some test cases were used to check whether task is completed or not.
- Converting the tags column from a string to an array of strings
- Joining Recipe Data to Review Data and Read the second data file
- Creating time-based features
- Saving the data we have created so far in a parquet file.

interaction_level_df.write.parquet('s3://bucket-bharath/data/interaction_level_df_processed.parquet')

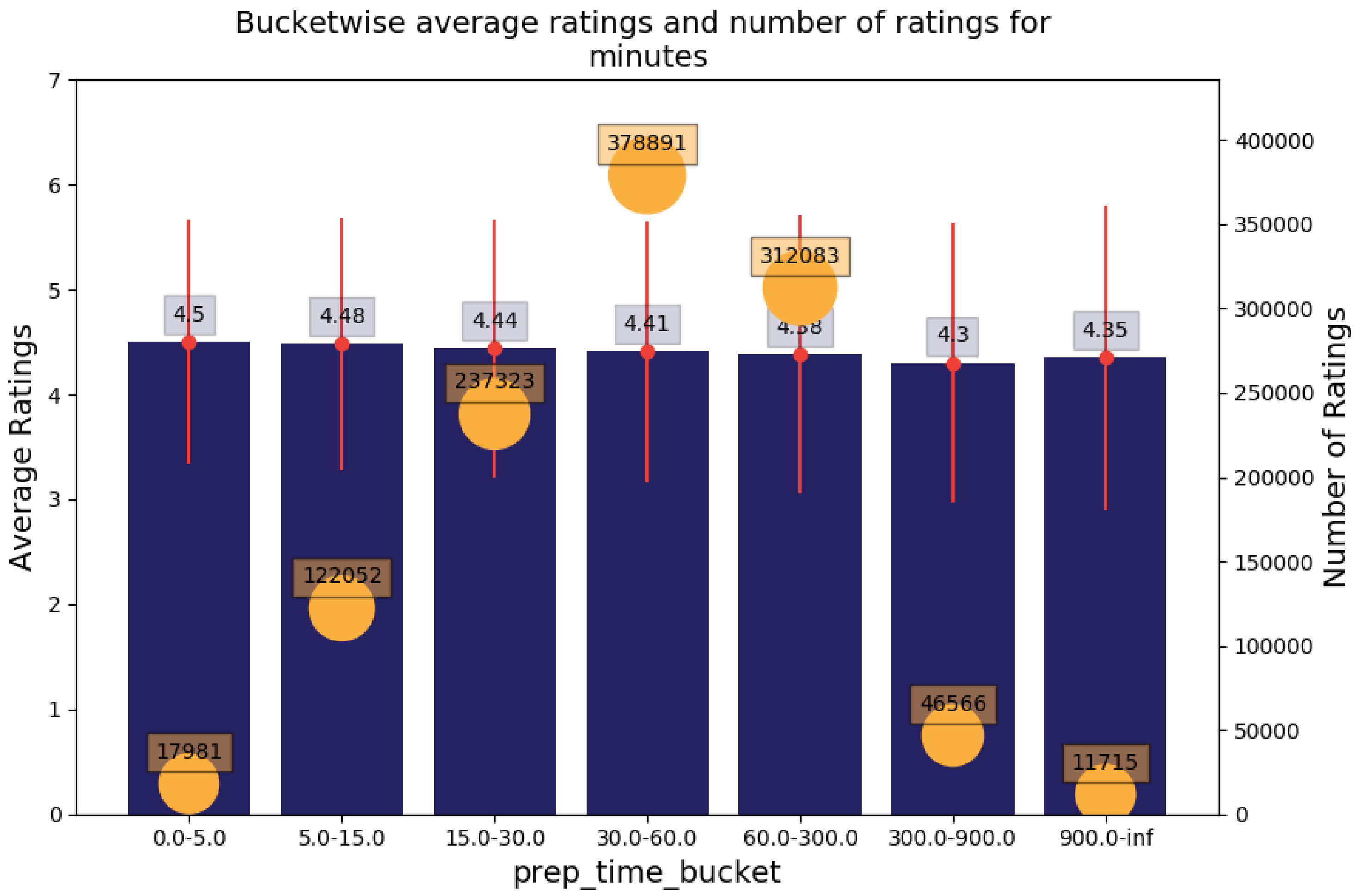
Exploratory Data Analysis
EDA

Bucketing and Cleaning Numerical Features



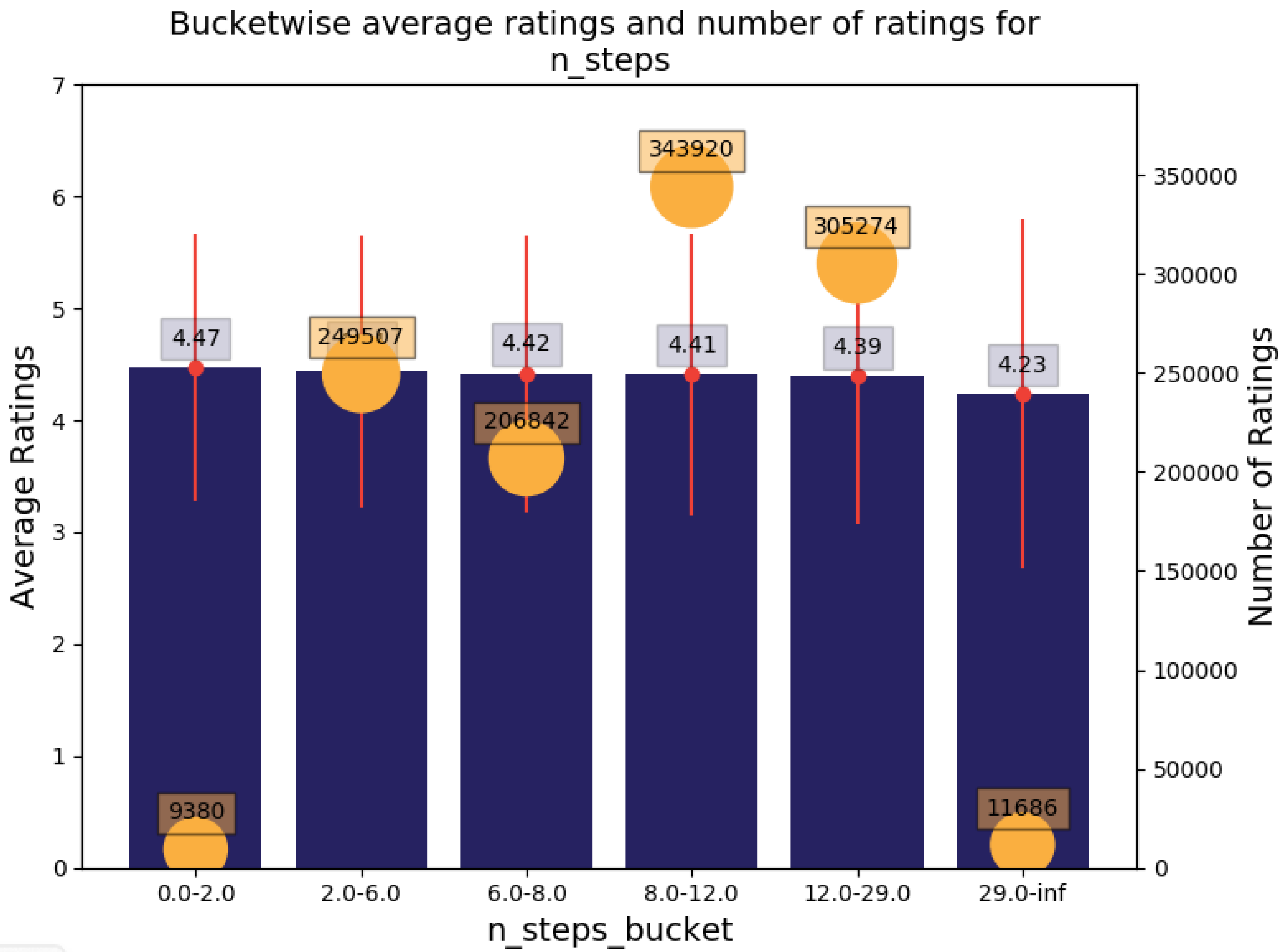
•**OBSERVATIONS**

•**Recipes more than 6 years old are rated low**



OBSERVATIONS

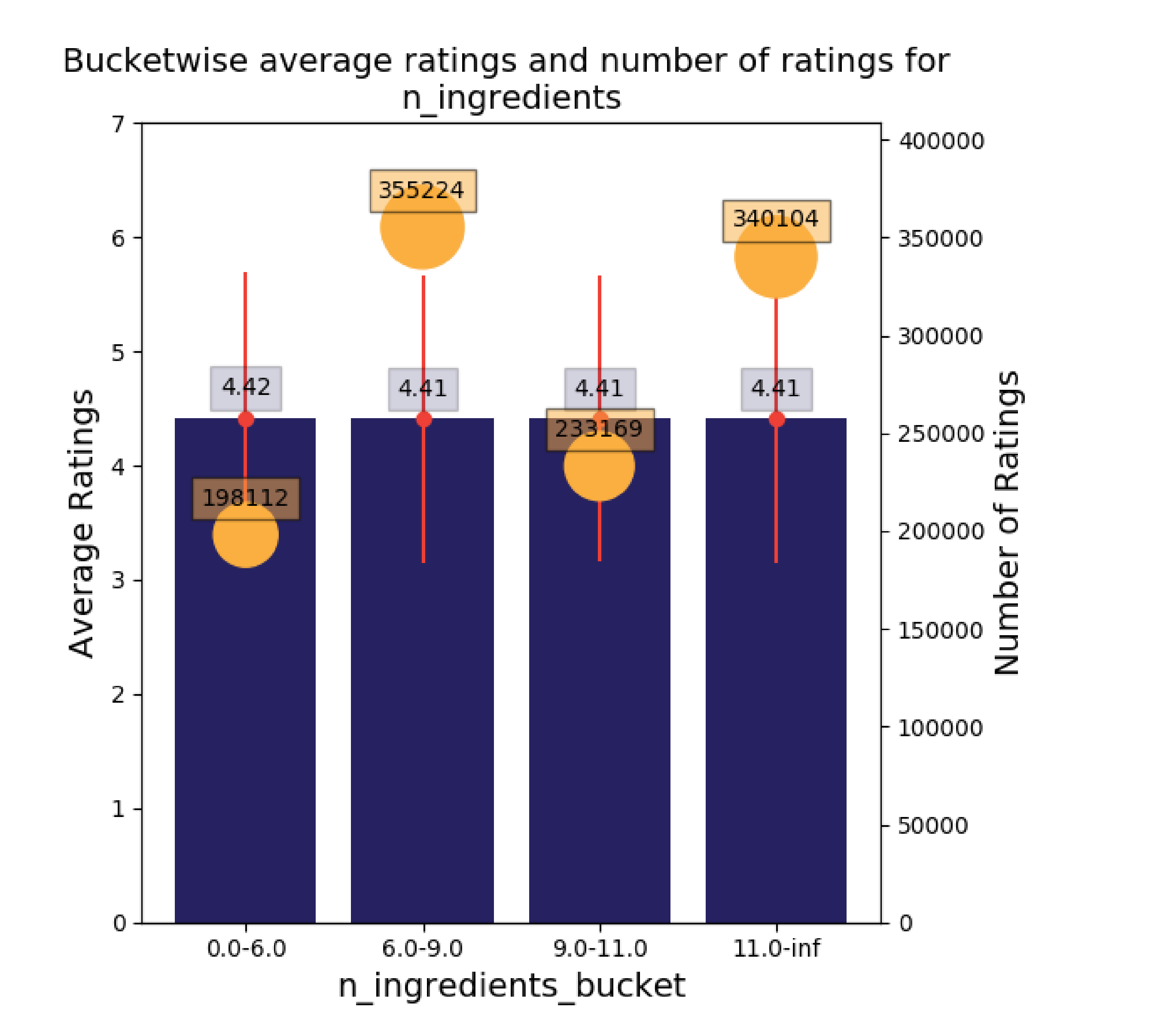
Minutes –
It is Somewhat relevant.
Low prep time is more preferred.



OBSERVATIONS

n_Steps-

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

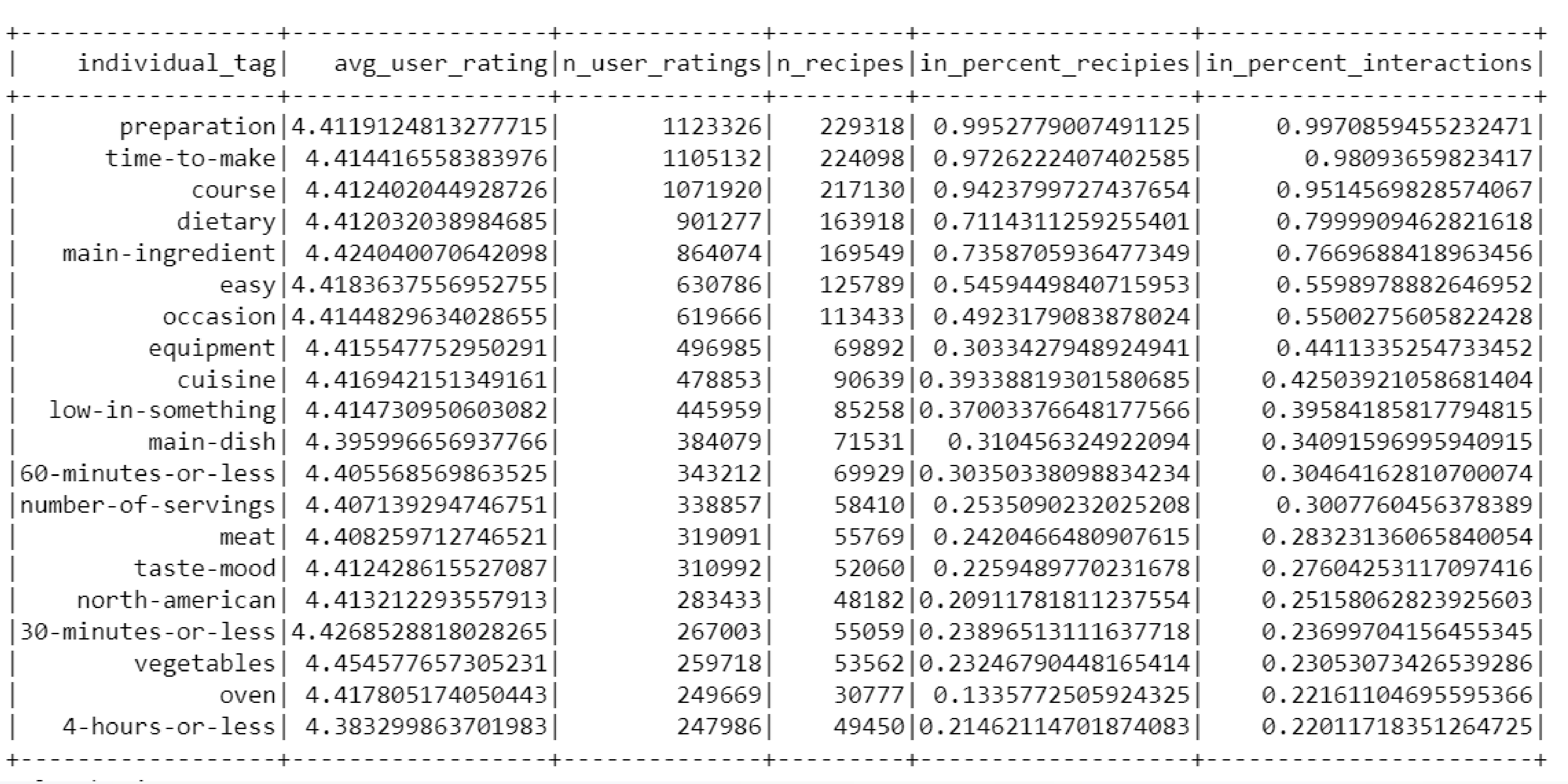


OBSERVATIONS

n_ingredients

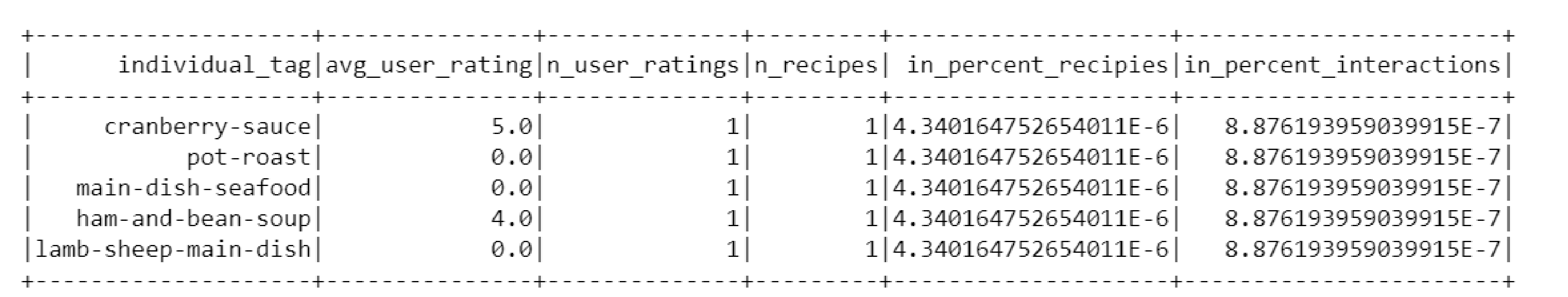
•Not relevant

Top 20 rated tags



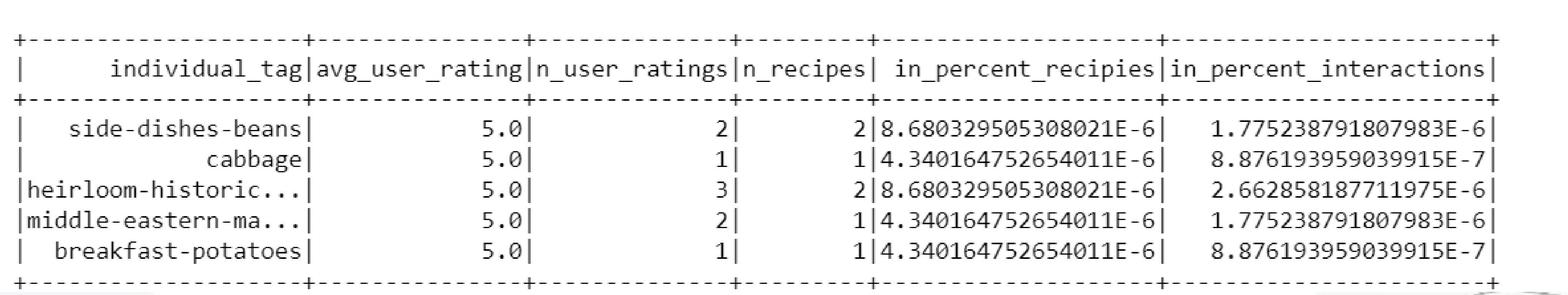
```
+------------------+------------------+--------------+---------+-------------------+-----------------------+
|    individual_tag|   avg_user_rating|n_user_ratings|n_recipes|in_percent_recipies|in_percent_interactions|
+------------------+------------------+--------------+---------+-------------------+-----------------------+
|       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|
+------------------+------------------+--------------+---------+-------------------+-----------------------+
```

Bottom five in tag_rating



```
+-------------------+--------------+-------------+---------+--------------------+----------------------+
|     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|
+-------------------+--------------+-------------+---------+--------------------+----------------------+
```

Top 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|
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
```

OBSERVATIONS - We can clearly observe Top 5 tags have low number of ratings.

Nutrition columns

- calories - Calories per serving seems irrelevant
- fat (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
- 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

More features

With rating = 5

1.User average years between review and submission high ratings .

2.User average Preparation time recipes reviewed high ratings

3.User average number of steps recipes reviewed high ratings

4. User average number of ingredients recipes reviewed high ratings

Conclusion And Recommendations

Top most features

a) Review time since submission
b) Number of steps
c) Preparation Times
d) Number of ingredients

- Number of ingredients in a recipe is not found to be relevant to the rating.
- The nutrition column such as fat, protein, sodium, sugar, calories are not found to be relevant in determining the rating of a recipe.
- Recipes reviewed by users after a long time from the submission date, having less number of steps, less number of ingredients and having less preparation time tend to have high ratings ie 5.

THANK YOU !!