Recipe Recommendation assignment :-

By Gaurav Chaudhary, Gaganpreet Kaur, Malaika Goveas (DSC- 59)

Problem Statement

Step into the shoes of an ML engineer working at food.com. Your job is to design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at.

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 your site reading about recipes.
Higher user engagement will likely result in more business opportunities like collaborations, promotions, etc.

The performance of a recommendation engine will significantly impact the revenue your recipe site can generate.

Designing a recommender from scratch is a time-consuming task.
In this assignment, you are expected to explore the data and create features that will be used to build the recommender.

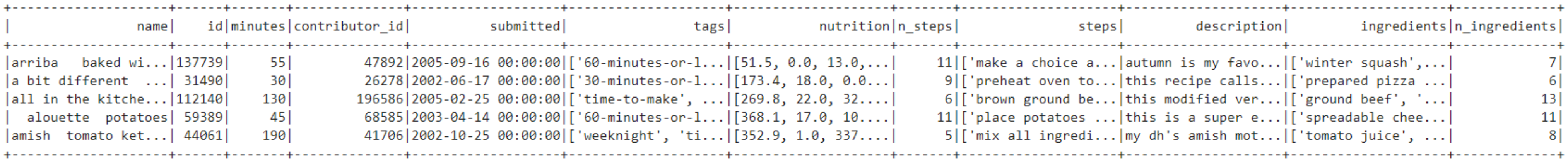
You will be working with the two CSV files linked below.

1. **Raw_recipes_cleaned.csv** - The first file is the Raw_recipes.csv file. It contains all the recipe-related information. Each row in this file describes a recipe
2. **RAW_interactions_cleaned.csv** - The second file we will be using is the RAW_interactions.csv. Each row in this data file is one user reviewing one recipe. One user can review more than one recipe, and each recipe can be reviewed by more than one user, so there is a many-to-many relationship between users and recipes, but the combination of user_id and reviewer_id in each row will be unique.

Task List

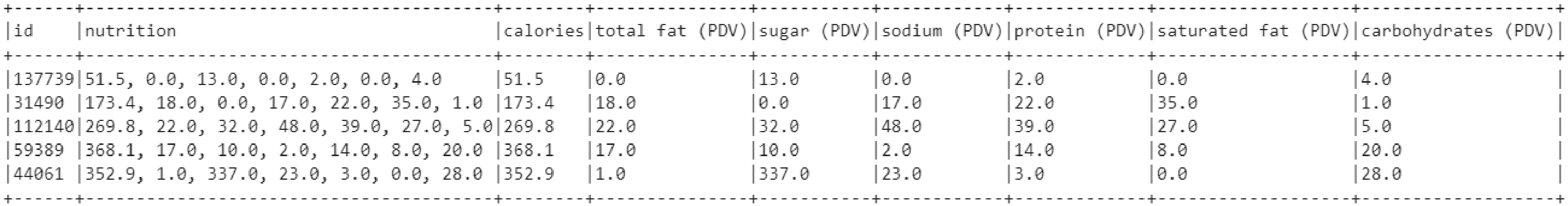
Task 1: Read the data

1. Read RAW_recipes.csv from S3 bucket.
2. Ensure each field has the correct data type.



```
+--------------------+------+-------+--------------+-------------------+--------------------+--------------------+-------+--------------------+--------------------+--------------------+-------------+
|                name|    id|minutes|contributor_id|          submitted|                tags|           nutrition|n_steps|               steps|         description|         ingredients|n_ingredients|
+--------------------+------+-------+--------------+-------------------+--------------------+--------------------+-------+--------------------+--------------------+--------------------+-------------+
|arriba   baked wi...|137739|     55|         47892|2005-09-16 00:00:00|['60-minutes-or-l...|[51.5, 0.0, 13.0,...|     11|['make a choice a...|autumn is my favo...|['winter squash',...|            7|
|a bit different  ...| 31490|     30|         26278|2002-06-17 00:00:00|['30-minutes-or-l...|[173.4, 18.0, 0.0...|      9|['preheat oven to...|this recipe calls...|['prepared pizza ...|            6|
|all in the kitche...|112140|    130|        196586|2005-02-25 00:00:00|['time-to-make', ...|[269.8, 22.0, 32....|      6|['brown ground be...|this modified ver...|['ground beef', '...|           13|
|  alouette  potatoes| 59389|     45|         68585|2003-04-14 00:00:00|['60-minutes-or-l...|[368.1, 17.0, 10....|     11|['place potatoes ...|this is a super e...|['spreadable chee...|           11|
|amish  tomato ket...| 44061|    190|         41706|2002-10-25 00:00:00|['weeknight', 'ti...|[352.9, 1.0, 337....|      5|['mix all ingredi...|my dh's amish mot...|['tomato juice', ...|            8|
+--------------------+------+-------+--------------+-------------------+--------------------+--------------------+-------+--------------------+--------------------+--------------------+-------------+
```

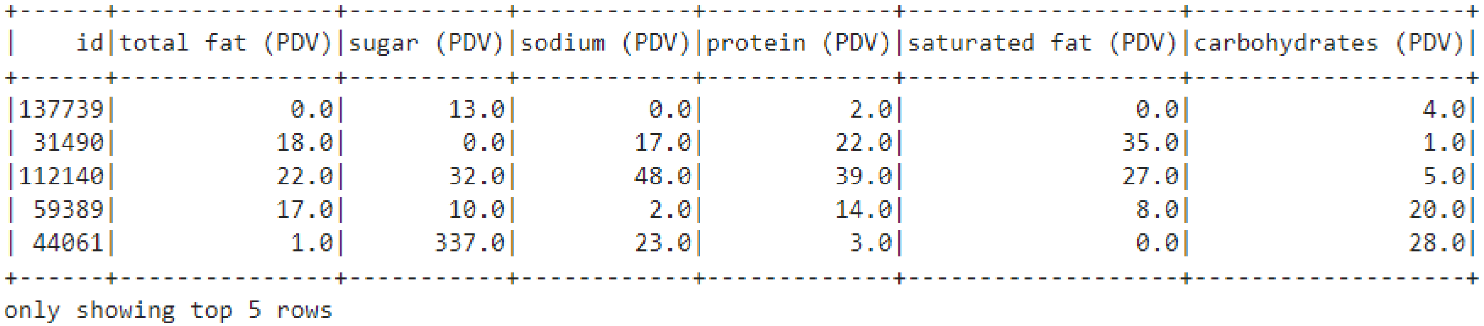
Task 2: Extract individual features from the nutrition column.



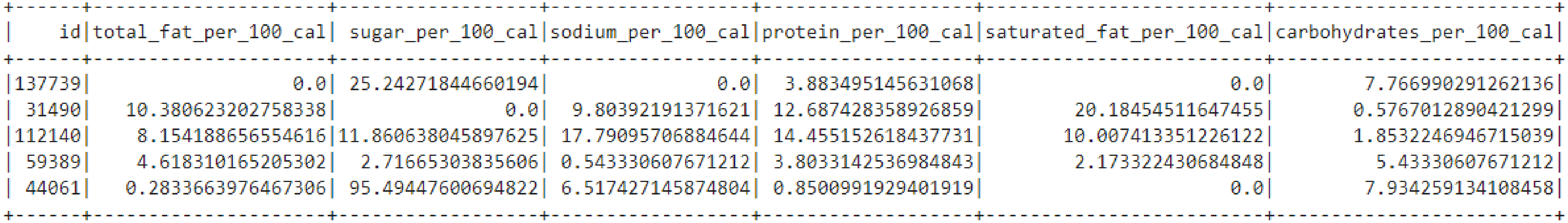
```
+------+----------------------------------------+--------+---------------+-----------+------------+-------------+-------------------+-------------------+
|id    |nutrition                               |calories|total fat (PDV)|sugar (PDV)|sodium (PDV)|protein (PDV)|saturated fat (PDV)|carbohydrates (PDV)|
+------+----------------------------------------+--------+---------------+-----------+------------+-------------+-------------------+-------------------+
|137739|51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0     |51.5    |0.0            |13.0       |0.0         |2.0          |0.0                |4.0                |
|31490 |173.4, 18.0, 0.0, 17.0, 22.0, 35.0, 1.0 |173.4   |18.0           |0.0        |17.0        |22.0         |35.0               |1.0                |
|112140|269.8, 22.0, 32.0, 48.0, 39.0, 27.0, 5.0|269.8   |22.0           |32.0       |48.0        |39.0         |27.0               |5.0                |
|59389 |368.1, 17.0, 10.0, 2.0, 14.0, 8.0, 20.0 |368.1   |17.0           |10.0       |2.0         |14.0         |8.0                |20.0               |
|44061 |352.9, 1.0, 337.0, 23.0, 3.0, 0.0, 28.0 |352.9   |1.0            |337.0      |23.0        |3.0          |0.0                |28.0               |
+------+----------------------------------------+--------+---------------+-----------+------------+-------------+-------------------+-------------------+
```

Task 3: Standardize the nutrition values.

Convert the nutritional values to per 100 calories.

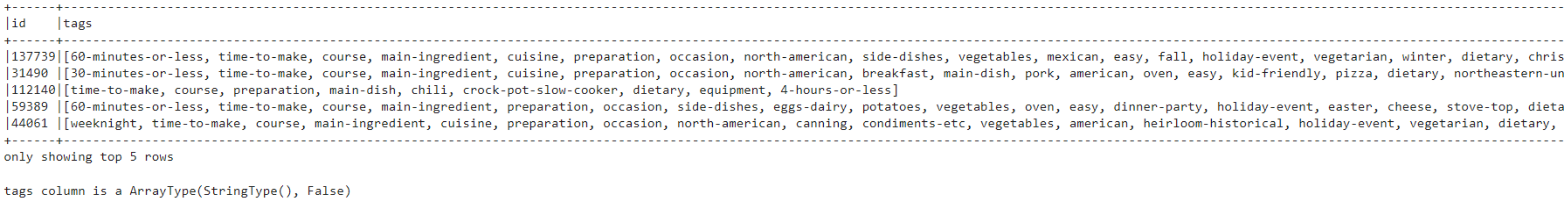


```
+------+---------------+-----------+------------+-------------+-------------------+-------------------+
|    id|total fat (PDV)|sugar (PDV)|sodium (PDV)|protein (PDV)|saturated fat (PDV)|carbohydrates (PDV)|
+------+---------------+-----------+------------+-------------+-------------------+-------------------+
|137739|            0.0|       13.0|         0.0|          2.0|                0.0|                4.0|
| 31490|           18.0|        0.0|        17.0|         22.0|               35.0|                1.0|
|112140|           22.0|       32.0|        48.0|         39.0|               27.0|                5.0|
| 59389|           17.0|       10.0|         2.0|         14.0|                8.0|               20.0|
| 44061|            1.0|      337.0|        23.0|          3.0|                0.0|               28.0|
+------+---------------+-----------+------------+-------------+-------------------+-------------------+
only showing top 5 rows
```



```
+------+---------------------+------------------+------------------+-------------------+-------------------------+-------------------------+
|    id|total_fat_per_100_cal| sugar_per_100_cal|sodium_per_100_cal|protein_per_100_cal|saturated_fat_per_100_cal|carbohydrates_per_100_cal|
+------+---------------------+------------------+------------------+-------------------+-------------------------+-------------------------+
|137739|                  0.0| 25.24271844660194|               0.0|  3.883495145631068|                      0.0|        7.766990291262136|
| 31490|   10.380623202758338|               0.0|  9.80392191371621| 12.687428358926859|        20.18454511647455|       0.5767012890421299|
|112140|    8.154188656554616|11.860638045897625| 17.79095706884644| 14.455152618437731|       10.007413351226122|        1.853224694715039|
| 59389|    4.618310165205302|  2.71665303835606| 0.543330607671212| 3.8033142536984843|        2.173322430684848|         5.43330607671212|
| 44061|   0.2833663976467306| 95.49447600694822| 6.517427145874804| 0.8500991929401919|                      0.0|        7.934259134108458|
+------+---------------------+------------------+------------------+-------------------+-------------------------+-------------------------+
```

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

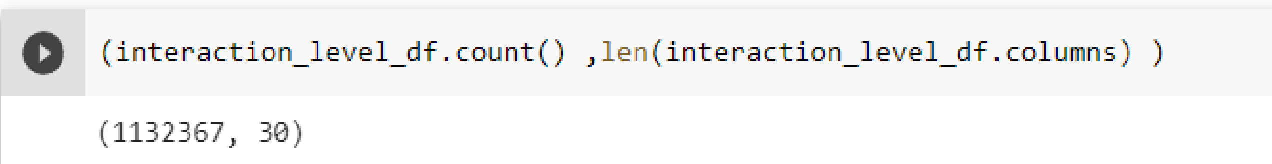


```
+------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|id    |tags
+------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|137739|[60-minutes-or-less, time-to-make, course, main-ingredient, cuisine, preparation, occasion, north-american, side-dishes, vegetables, mexican, easy, fall, holiday-event, vegetarian, winter, dietary, chris
|31490 |[30-minutes-or-less, time-to-make, course, main-ingredient, cuisine, preparation, occasion, north-american, breakfast, main-dish, pork, american, oven, easy, kid-friendly, pizza, dietary, northeastern-un
|112140|[time-to-make, course, preparation, main-dish, chili, crock-pot-slow-cooker, dietary, equipment, 4-hours-or-less]
|59389 |[60-minutes-or-less, time-to-make, course, main-ingredient, preparation, occasion, side-dishes, eggs-dairy, potatoes, vegetables, oven, easy, dinner-party, holiday-event, easter, cheese, stove-top, dieta
|44061 |[weeknight, time-to-make, course, main-ingredient, cuisine, preparation, occasion, north-american, canning, condiments-etc, vegetables, american, heirloom-historical, holiday-event, vegetarian, dietary,
+------+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
only showing top 5 rows

tags column is a ArrayType(StringType(), False)
```

Task 5: Read the second data file

Read the RAW_interaction.csv and join this interaction level file with the recipe level data frame.

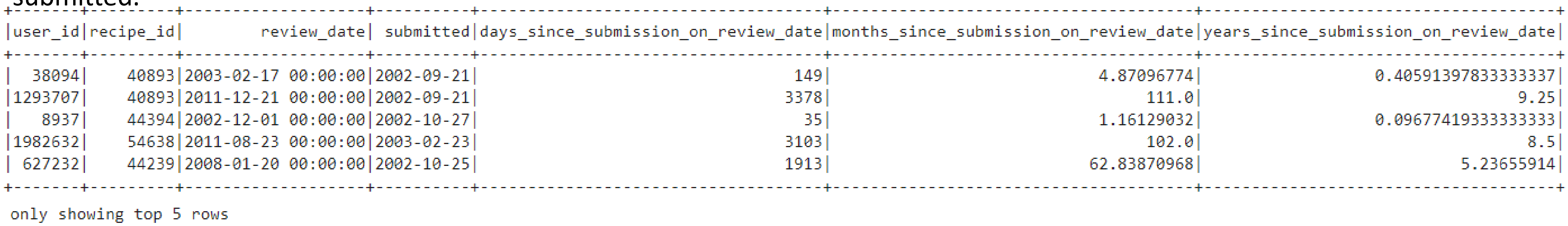


```
(interaction_level_df.count() ,len(interaction_level_df.columns) )
```

```
(1132367, 30)
```

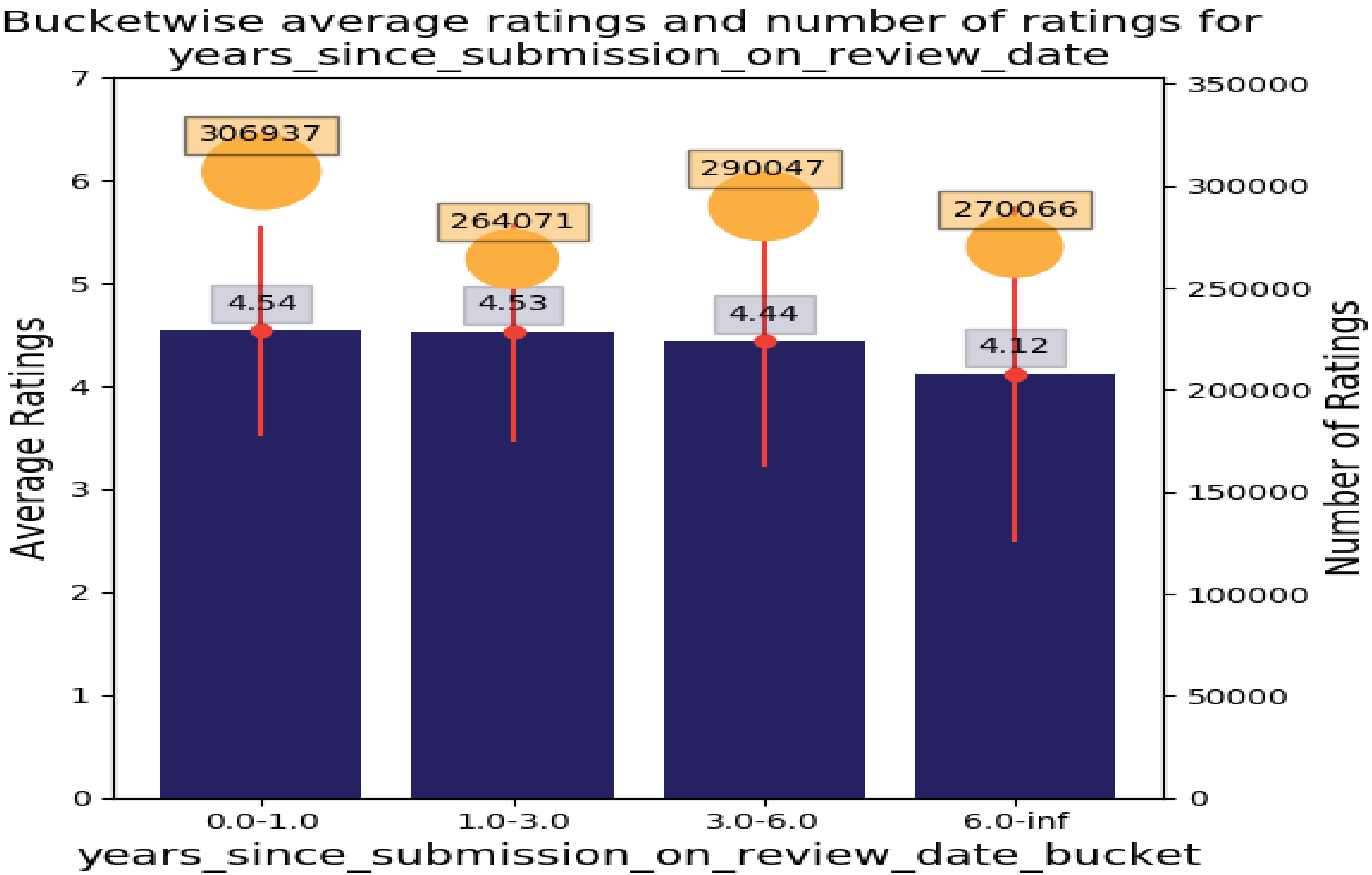
Task 6: Create time-based features.

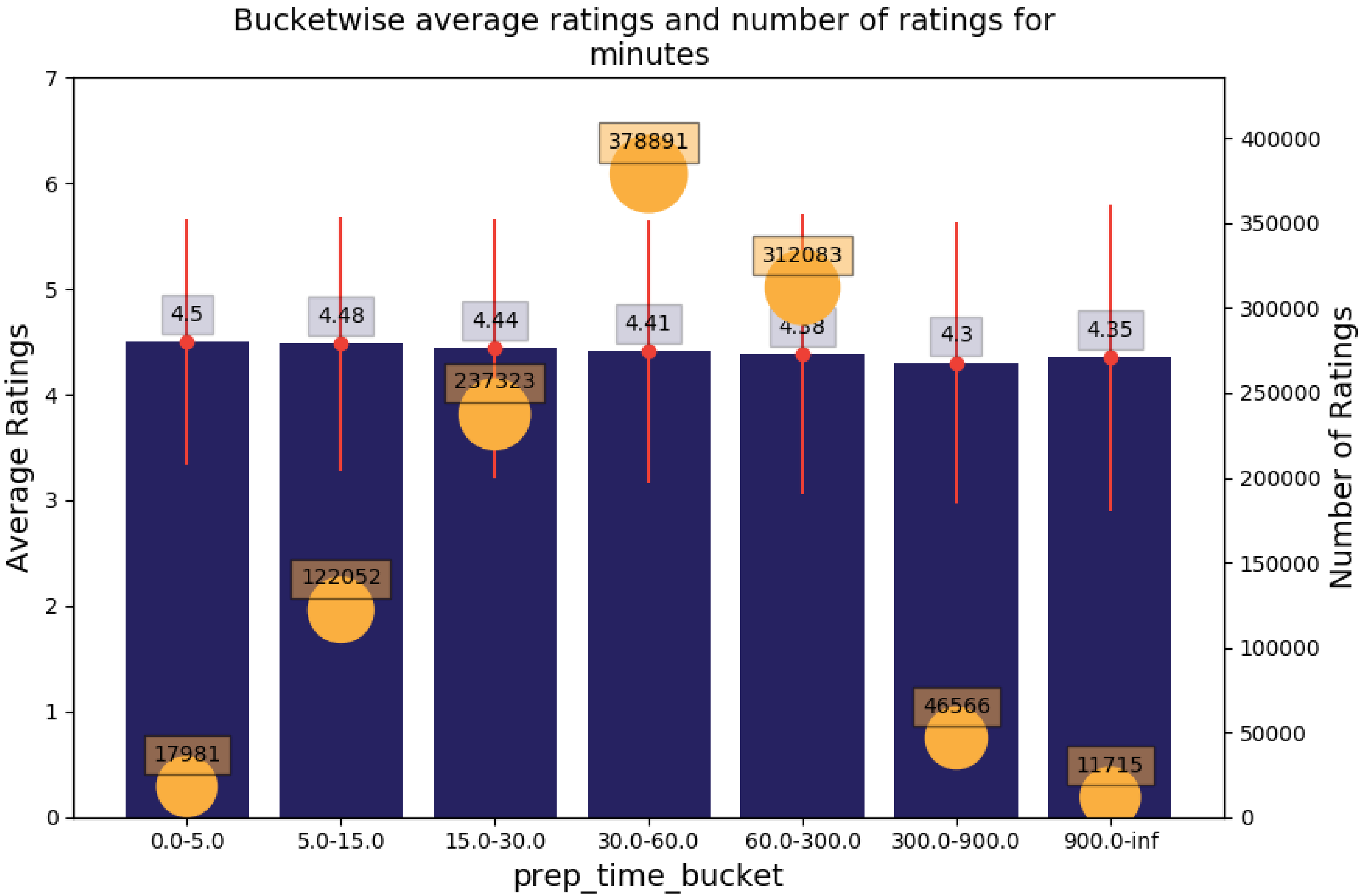
Create features that capture the time passed between one review and the date on which the recipe was submitted.



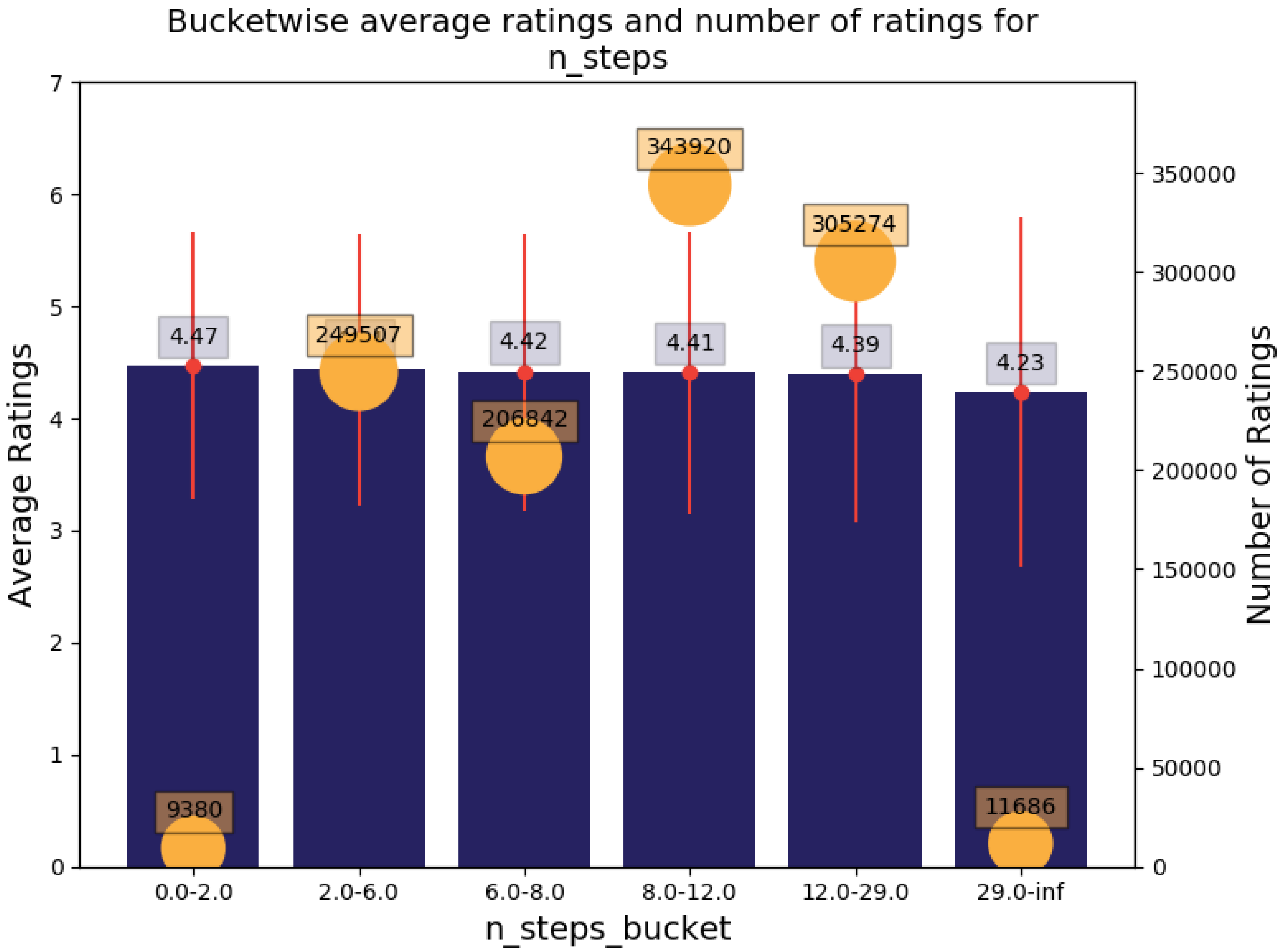
```
+-------+---------+-------------------+----------+------------------------------------+--------------------------------------+-------------------------------------+
|user_id|recipe_id|        review_date| submitted|days_since_submission_on_review_date|months_since_submission_on_review_date|years_since_submission_on_review_date|
+-------+---------+-------------------+----------+------------------------------------+--------------------------------------+-------------------------------------+
|  38094|    40893|2003-02-17 00:00:00|2002-09-21|                                 149|                            4.87096774|                  0.40591397833333337|
|1293707|    40893|2011-12-21 00:00:00|2002-09-21|                                3378|                                 111.0|                                 9.25|
|   8937|    44394|2002-12-01 00:00:00|2002-10-27|                                  35|                            1.16129032|                  0.09677419333333333|
|1982632|    54638|2011-08-23 00:00:00|2003-02-23|                                3103|                                 102.0|                                  8.5|
| 627232|    44239|2008-01-20 00:00:00|2002-10-25|                                1913|                           62.83870968|                           5.23655914|
+-------+---------+-------------------+----------+------------------------------------+--------------------------------------+-------------------------------------+
only showing top 5 rows
```

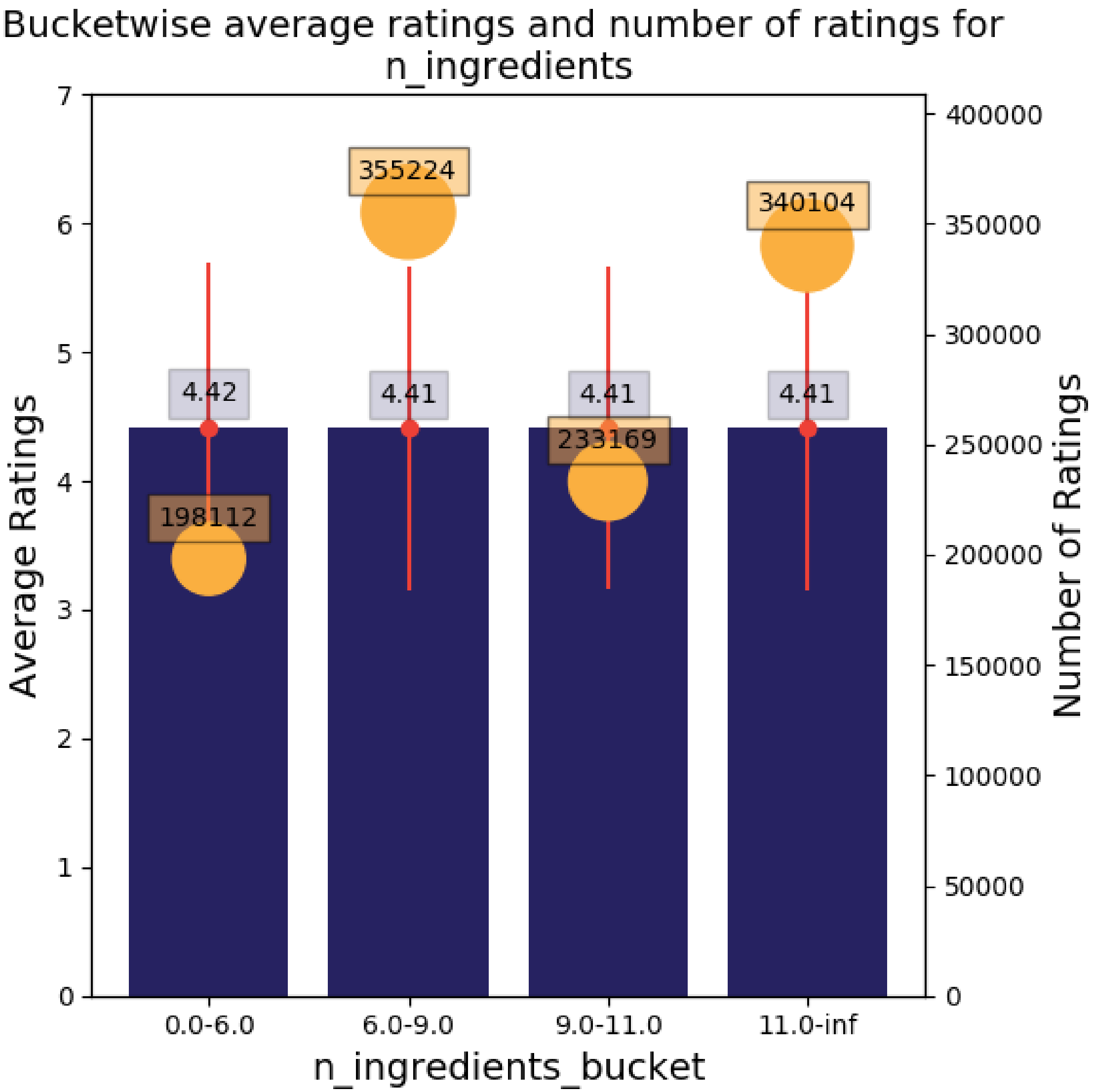
Task 7: Processing Numerical Columns & Exploratory Data Analysis

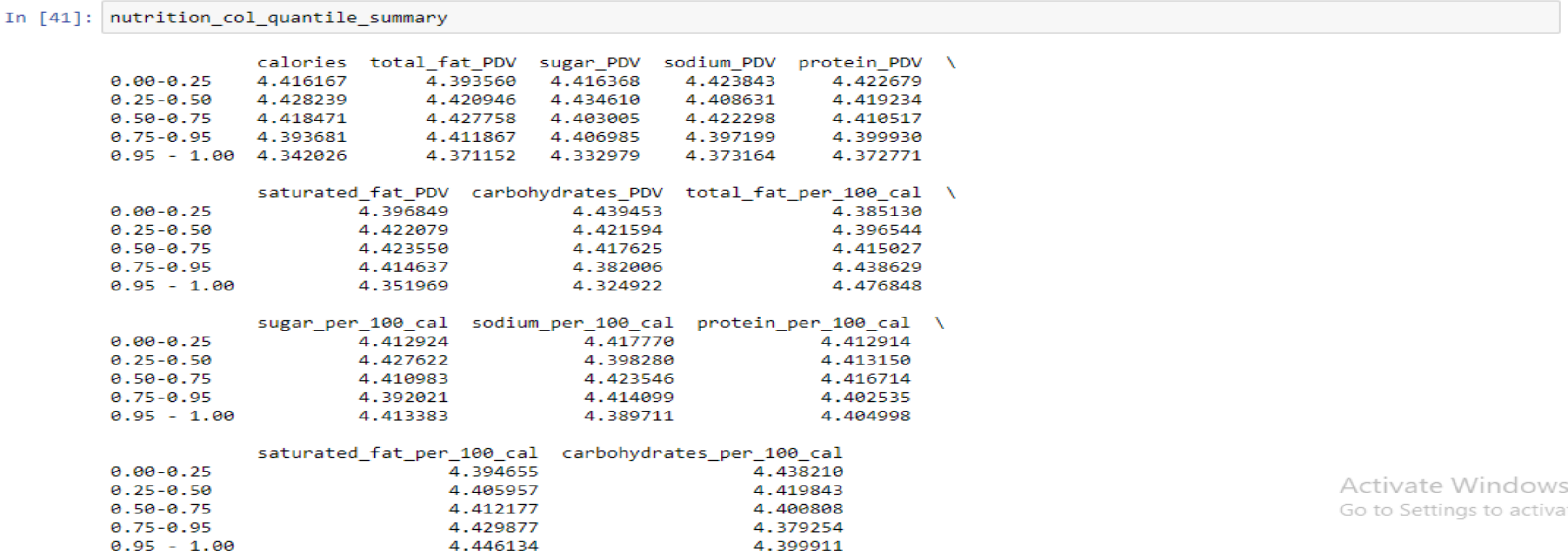




Bucketwise average ratings and number of ratings for minutes


Bucketwise average ratings and number of ratings for n_steps
Average Ratings
0
1
2
3
4
5
6
7
Number of Ratings
0
50000
100000
150000
200000
250000
300000
350000
n_steps_bucket
0.0-2.0
2.0-6.0
6.0-8.0
8.0-12.0
12.0-29.0
29.0-inf
4.47
4.42
4.41
4.39
4.23
9380
249507
206842
343920
305274
11686


Bucketwise average ratings and number of ratings for n_ingredients
Average Ratings
Number of Ratings
n_ingredients_bucket
0.0-6.0
6.0-9.0
9.0-11.0
11.0-inf
0
1
2
3
4
5
6
7
0
50000
100000
150000
200000
250000
300000
350000
400000
4.42
4.41
4.41
4.41
198112
355224
233169
340104



```
In [41]: nutrition_col_quantile_summary
```

```
             calories  total_fat_PDV  sugar_PDV  sodium_PDV  protein_PDV  \
0.00-0.25    4.416167       4.393560   4.416368    4.423843     4.422679
0.25-0.50    4.428239       4.420946   4.434610    4.408631     4.419234
0.50-0.75    4.418471       4.427758   4.403005    4.422298     4.410517
0.75-0.95    4.393681       4.411867   4.406985    4.397199     4.399930
0.95 - 1.00  4.342026       4.371152   4.332979    4.373164     4.372771

             saturated_fat_PDV  carbohydrates_PDV  total_fat_per_100_cal  \
0.00-0.25             4.396849           4.439453               4.385130
0.25-0.50             4.422079           4.421594               4.396544
0.50-0.75             4.423550           4.417625               4.415027
0.75-0.95             4.414637           4.382006               4.438629
0.95 - 1.00           4.351969           4.324922               4.476848

             sugar_per_100_cal  sodium_per_100_cal  protein_per_100_cal  \
0.00-0.25             4.412924            4.417770             4.412914
0.25-0.50             4.427622            4.398280             4.413150
0.50-0.75             4.410983            4.423546             4.416714
0.75-0.95             4.392021            4.414099             4.402535
0.95 - 1.00           4.413383            4.389711             4.404998

             saturated_fat_per_100_cal  carbohydrates_per_100_cal
0.00-0.25                     4.394655                   4.438210
0.25-0.50                     4.405957                   4.419843
0.50-0.75                     4.412177                   4.400808
0.75-0.95                     4.429877                   4.379254
0.95 - 1.00                   4.446134                   4.399911
```

Columns to be bucketized.

1. years_since_submission_on_review_date
2. minutes
3. calories
4. total_fat_PDV
5. sugar_PDV
6. sodium_PDV
7. protein_PDV
8. saturated_fat_PDV
9. carbohydrates_PDV

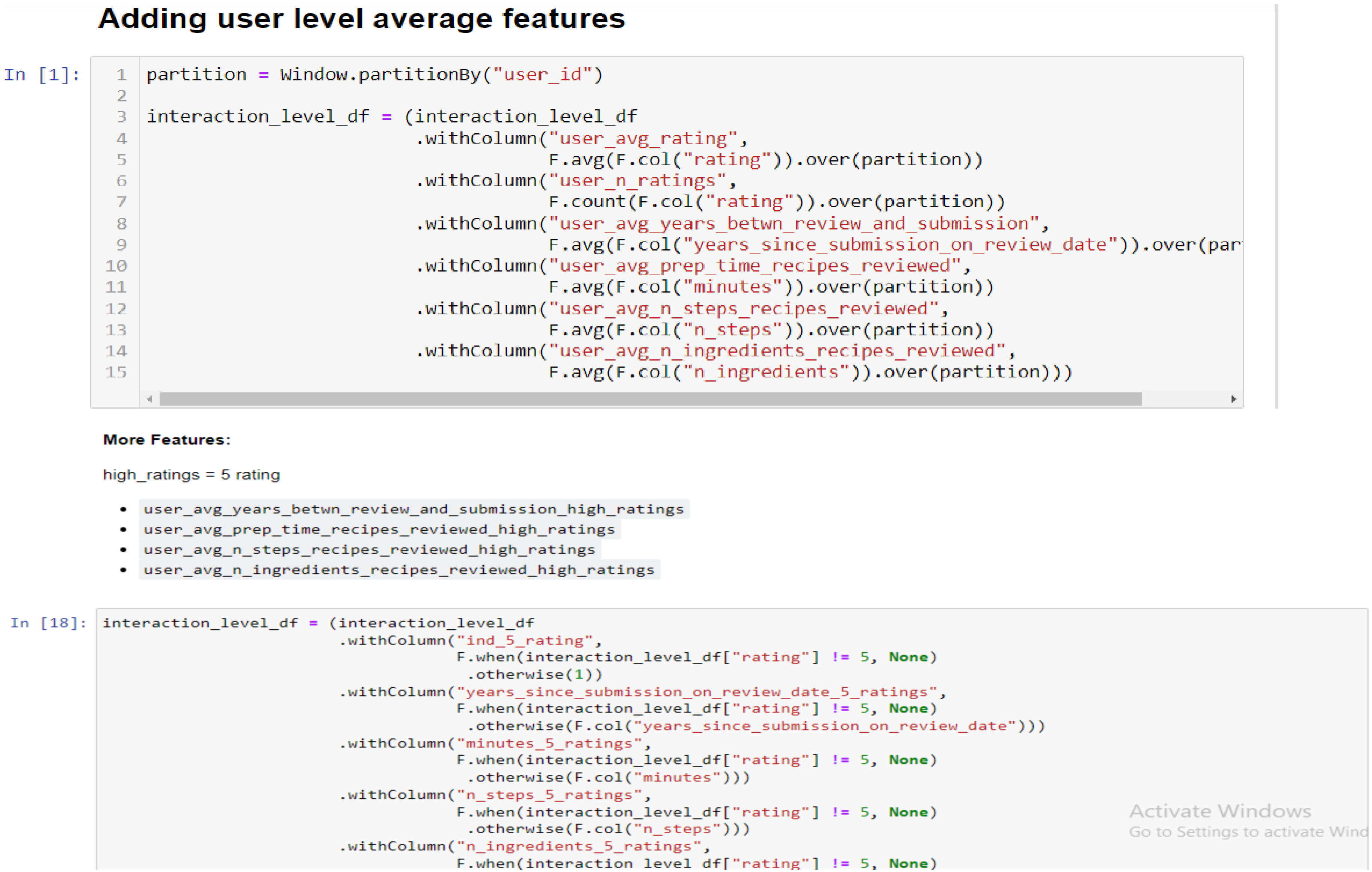
After creating buckets, study the variation of the average rating for each bucket and decide whether or not a particular bucketed column should be kept in the analysis.

Task 8: Create user-level features

create the following user-level features:

- user_avg_rating
- user_avg_n_ratings
- user_avg_years_betwn_review_and_submission
- user_avg_prep_time_recipes_reviewed
- user_avg_n_steps_recipes_reviewed
- user_avg_n_ingredients_recipes_reviewed
- user_avg_years_betwn_review_and_submission_high_ratings
- user_avg_calories_recipes_reviewed
- user_avg_total_fat_per_100_cal_recipes_reviewed
- user_avg_sugar_per_100_cal_recipes_reviewed
- user_avg_sodium_per_100_cal_recipes_reviewed
- user_avg_protein_per_100_cal_recipes_reviewed
- user_avg_saturated_fat_per_100_cal_recipes_reviewed
- user_avg_carbohydrates_per_100_cal_recipes_reviewed
- user_avg_prep_time_recipes_reviewed_high_ratings
- user_avg_n_steps_recipes_reviewed_high_ratings
- user_avg_n_ingredients_recipes_reviewed_high_ratings

Here, high ratings refer to only those reviews where the user has given five ratings to a recipe.



Adding user level average features

In [1]:
```
partition = Window.partitionBy("user_id")

interaction_level_df = (interaction_level_df
                        .withColumn("user_avg_rating",
                                    F.avg(F.col("rating")).over(partition))
                        .withColumn("user_n_ratings",
                                    F.count(F.col("rating")).over(partition))
                        .withColumn("user_avg_years_betwn_review_and_submission",
                                    F.avg(F.col("years_since_submission_on_review_date")).over(par
                        .withColumn("user_avg_prep_time_recipes_reviewed",
                                    F.avg(F.col("minutes")).over(partition))
                        .withColumn("user_avg_n_steps_recipes_reviewed",
                                    F.avg(F.col("n_steps")).over(partition))
                        .withColumn("user_avg_n_ingredients_recipes_reviewed",
                                    F.avg(F.col("n_ingredients")).over(partition)))
```

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`

In [18]:
```
interaction_level_df = (interaction_level_df
                        .withColumn("ind_5_rating",
                                    F.when(interaction_level_df["rating"] != 5, None)
                                     .otherwise(1))
                        .withColumn("years_since_submission_on_review_date_5_ratings",
                                    F.when(interaction_level_df["rating"] != 5, None)
                                     .otherwise(F.col("years_since_submission_on_review_date")))
                        .withColumn("minutes_5_ratings",
                                    F.when(interaction_level_df["rating"] != 5, None)
                                     .otherwise(F.col("minutes")))
                        .withColumn("n_steps_5_ratings",
                                    F.when(interaction_level_df["rating"] != 5, None)
                                     .otherwise(F.col("n_steps")))
                        .withColumn("n_ingredients_5_ratings",
                                    F.when(interaction_level_df["rating"] != 5, None)
```

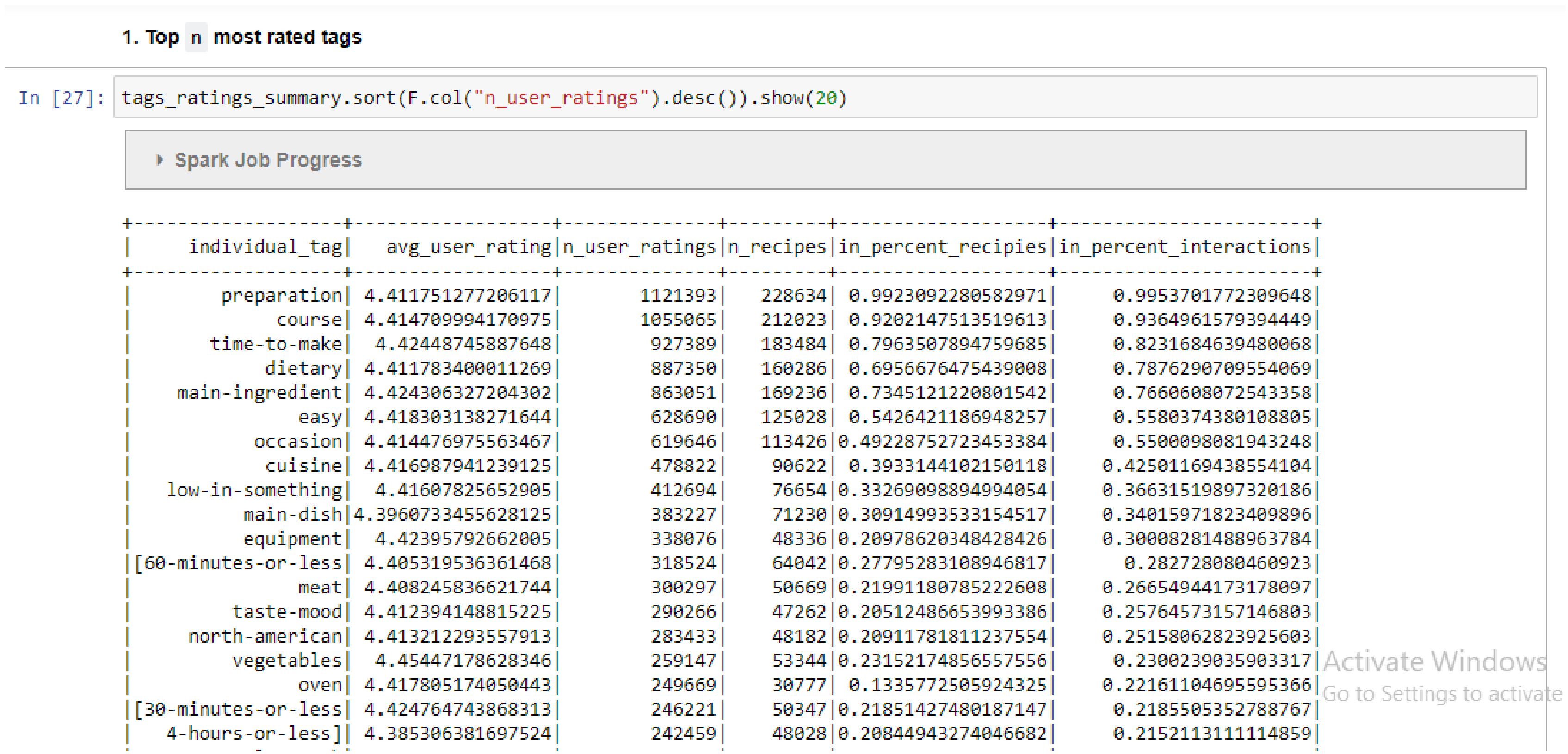
Task 9: Create tag-level features

Extract tags-level features. If you extract and list unique tags and explore all the available tags, you will realize that tags hold a lot of information about the recipe.

For example, the healthy tag signifies that the person who uploaded the recipe considers it healthy.

If a user specifically looks for the healthy tag, you would want to recommend more healthy recipes to them.

Find the most value-adding tags and create features to capture them.



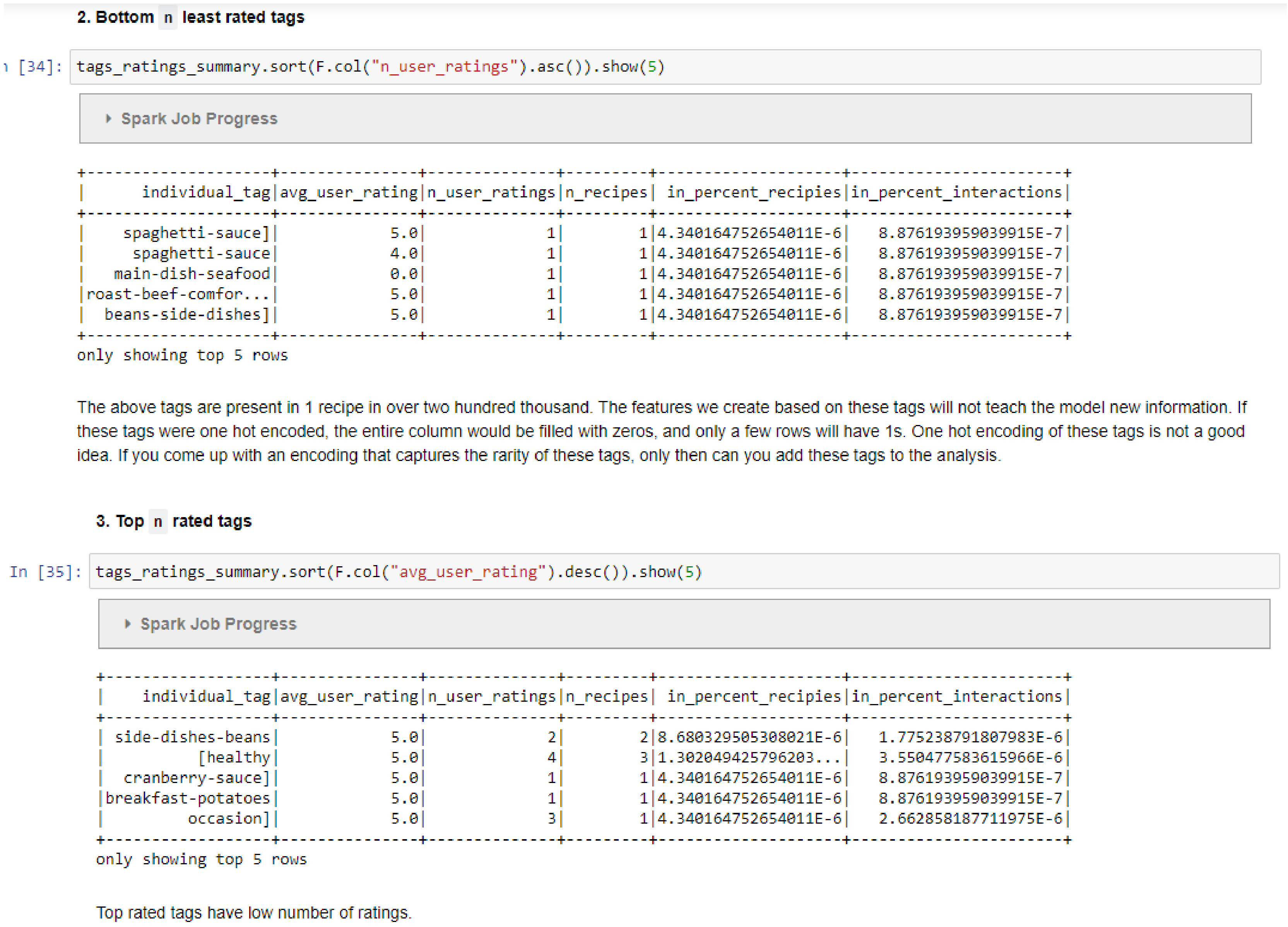
1. Top n most rated tags

In [27]:

```
tags_ratings_summary.sort(F.col("n_user_ratings").desc()).show(20)
```

▸ Spark Job Progress

```
+------------------+-----------------+--------------+---------+-------------------+----------------------+
|     individual_tag|   avg_user_rating|n_user_ratings|n_recipes|in_percent_recipies|in_percent_interactions|
+------------------+-----------------+--------------+---------+-------------------+----------------------+
|        preparation| 4.411751277206117|       1121393|   228634| 0.9923092280582971|     0.9953701772309648|
|             course| 4.414709994170975|       1055065|   212023| 0.9202147513519613|     0.9364961579394449|
|       time-to-make|  4.42448745887648|        927389|   183484| 0.7963507894759685|     0.8231684639480068|
|            dietary| 4.411783400011269|        887350|   160286| 0.6956676475439008|     0.7876290709554069|
|    main-ingredient| 4.424306327204302|        863051|   169236| 0.7345121220801542|     0.7660608072543358|
|               easy| 4.418303138271644|        628690|   125028| 0.5426421186948257|     0.5580374380108805|
|           occasion| 4.414476975563467|        619646|   113426|0.49228752723453384|     0.5500098081943248|
|            cuisine| 4.416987941239125|        478822|    90622| 0.3933144102150118|    0.42501169438554104|
|   low-in-something|  4.41607825652905|        412694|    76654|0.33269098894994054|    0.36631519897320186|
|          main-dish|4.3960733455628125|        383227|    71230|0.30914993533154517|    0.34015971823409896|
|          equipment|  4.42395792662005|        338076|    48336|0.20978620348428426|    0.30008281488963784|
|[60-minutes-or-less| 4.405319536361468|        318524|    64042|0.27795283108946817|      0.282728080460923|
|               meat| 4.408245836621744|        300297|    50669|0.21991180785222608|    0.26654944173178097|
|         taste-mood| 4.412394148815225|        290266|    47262|0.20512486653993386|    0.25764573157146803|
|     north-american| 4.413212293557913|        283433|    48182|0.20911781811237554|    0.25158062823925603|
|         vegetables|  4.45447178628346|        259147|    53344|0.23152174856557556|     0.2300239035903317|
|               oven| 4.417805174050443|        249669|    30777| 0.1335772505924325|     0.22161104695595366|
|[30-minutes-or-less| 4.424764743868313|        246221|    50347|0.21851427480187147|     0.2185505352788767|
|    4-hours-or-less]| 4.385306381697524|        242459|    48028|0.20844943274046682|     0.2152113111114859|
```



2. Bottom `n` least rated tags

[34]:

```python
tags_ratings_summary.sort(F.col("n_user_ratings").asc()).show(5)
```

```
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
|      individual_tag|avg_user_rating|n_user_ratings|n_recipes| in_percent_recipies|in_percent_interactions|
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
|    spaghetti-sauce]|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|     spaghetti-sauce|            4.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|   main-dish-seafood|            0.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|roast-beef-comfor...|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|  beans-side-dishes]|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
+--------------------+---------------+--------------+---------+--------------------+-----------------------+
only showing top 5 rows
```

The above tags are present in 1 recipe in over two hundred thousand. The features we create based on these tags will not teach the model new information. If these tags were one hot encoded, the entire column would be filled with zeros, and only a few rows will have 1s. One hot encoding of these tags is not a good idea. If you come up with an encoding that captures the rarity of these tags, only then can you add these tags to the analysis.

3. Top `n` rated tags

In [35]:

```python
tags_ratings_summary.sort(F.col("avg_user_rating").desc()).show(5)
```

```
+------------------+---------------+--------------+---------+--------------------+-----------------------+
|    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|
|          [healthy|            5.0|             4|        3|1.302049425796203...|   3.550477583615966E-6|
|  cranberry-sauce]|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|breakfast-potatoes|            5.0|             1|        1|4.340164752654011E-6|   8.876193959039915E-7|
|         occasion]|            5.0|             3|        1|4.340164752654011E-6|   2.662858187711975E-6|
+------------------+---------------+--------------+---------+--------------------+-----------------------+
only showing top 5 rows
```

Top rated tags have low number of ratings.