

| **PROCESSING BIG DATASET**  **Exploratory Data Analysis – Receipt Datasets** |
| --- |

| **Developed by**  **Members:**   | **No.** | **Student Name** | **Student ID** | | --- | --- | --- | | 1 | Nguyen Huynh Ngoc Han (Jack) | Student1503324 | | 2 | Vu Manh Trung Hai | Student1522966 |  * **Class No.: DT2307L-NK** * **Start Date: 21st October 2024.** * **End Date: 25th November 2024.** * **Name of the Coordinator: Teacher Mr. Ho Nhat Minh** * **Date of Submission: 04th December 2024.** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

| Exploratory Data Analysis – Receipt Datasets |
| --- |

***This final assignment includes 3 sections as below:***

1. Assignment Overview & Objectives
2. Assignment Approach
3. 9 Tasks & Solutions

**I. ASSIGNMENT OVERVIEW & OBJECTIVE**

* + **Problem Statement:** Imagine you're an ML engineer at food.com. Your task is to build a recommendation engine that suggests recipes to users based on their current selection and past interactions. This system aims to increase user engagement and time spent on the website, leading to more business opportunities. The recommender system's performance directly impacts the website's revenue generation potential. This assignment focuses on the crucial step of EDA and feature extraction to prepare the data for building the recommender system.
* **Key objective:** Perform Exploratory Data Analysis (EDA) and feature extraction from raw recipe data to design a recommender system. Total 6 required tasks and 3 optional tasks.
* **Regarding the Data Source: 2 datasets as below**
* **1 - RAW\_recipes\_cleaned.csv:** Contains recipe-related information.

Download Link: <https://raw-recipes-clean-upgrad.s3.amazonaws.com/RAW_recipes_cleaned.csv>

Fields:

* name: Recipe name (String)
* id: Recipe ID (Integer)
* minutes: Preparation time in minutes (Integer)
* contributor\_id: User ID who submitted the recipe (Integer)
* submitted: Date of submission (Date)
* tags: Food.com tags for the recipe (Array of strings)
* nutrition: Nutritional information [calories, total fat (PDV), sugar (PDV), sodium (PDV), protein (PDV), saturated fat (PDV), carbohydrates (PDV)] (Array of floats)
* n\_steps: Number of steps in the recipe (Integer)
* steps: Recipe steps in order (String)
* description: Recipe description (String)
* ingredients: List of ingredients (Array of strings)
* n\_ingredients: Number of ingredients (Integer)
* **2 - RAW\_interactions\_cleaned.csv:** Contains user-recipe interaction data.

Download Link: <https://raw-interactions-upgrad.s3.amazonaws.com/RAW_interactions_cleaned.csv>

Fields:

* user\_id: User ID (Integer)
* recipe\_id: Recipe ID (Integer)
* date: Date of interaction (Date)
* rating: Rating given (Float)
* review: Review text (String)

*(\*) Link for final assignment here:*

[*https://drive.google.com/file/d/1cT6Xudv9xOKKOnXmO\_QvEuyiePB5g61C/view?usp=sharing*](https://drive.google.com/file/d/1cT6Xudv9xOKKOnXmO_QvEuyiePB5g61C/view?usp=sharing)

**II. ASSIGNMENT APPROACH**

The core approach for this assignment is to execute the nine specified tasks by leveraging the Databricks platform for Exploratory Data Analysis (EDA) and related analyses. This involves importing and transforming the two datasets into Python dataframes, followed by employing Python algorithms to systematically address the requirements of the tasks.

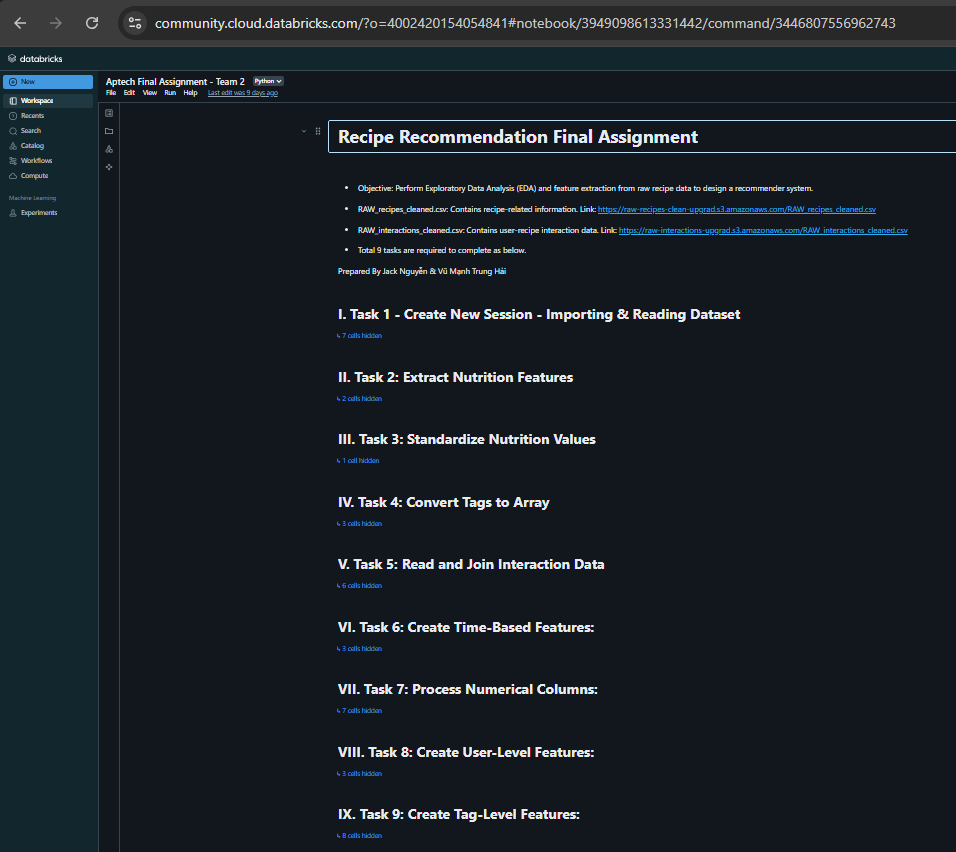
**III. 9 TASKS AND SOLUTIONS**

Here is the published final link from Databricks containing the implementation of all 9 tasks:

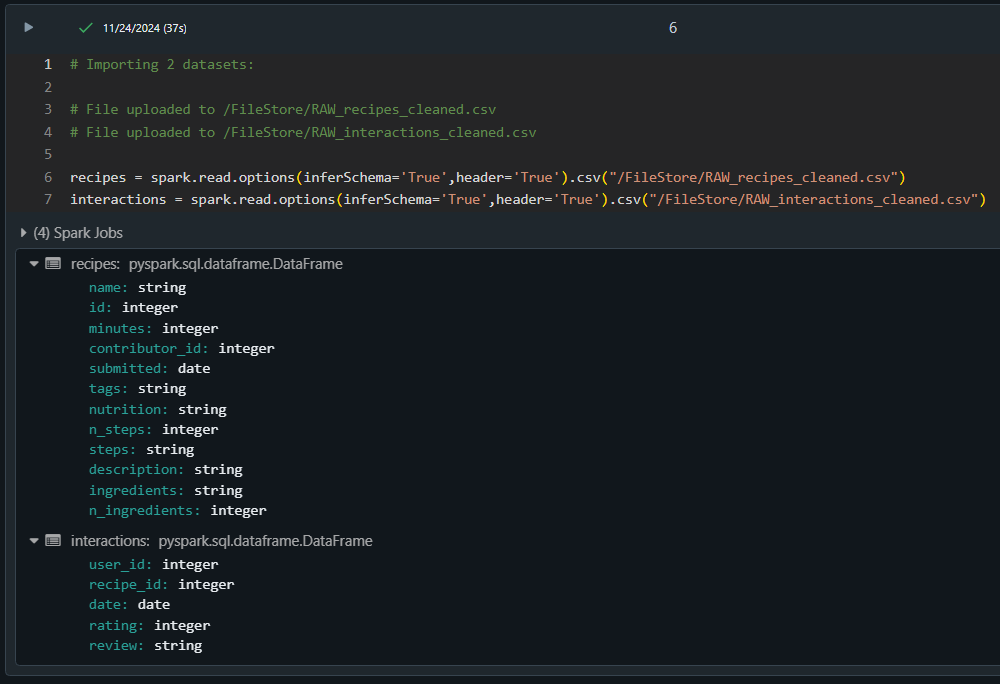
<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4002420154054841/3949098613331442/3655900734631075/latest.html>

Here is the github link as requested:

[LINK ADDED HERE]



* **Task 1: Read the datasets**

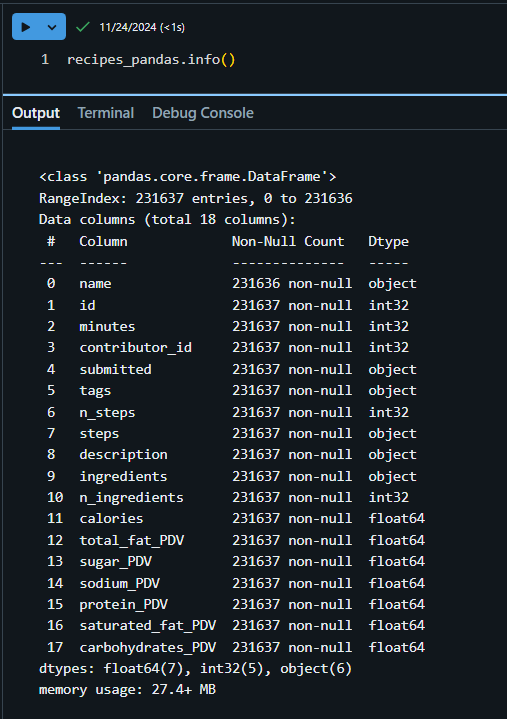


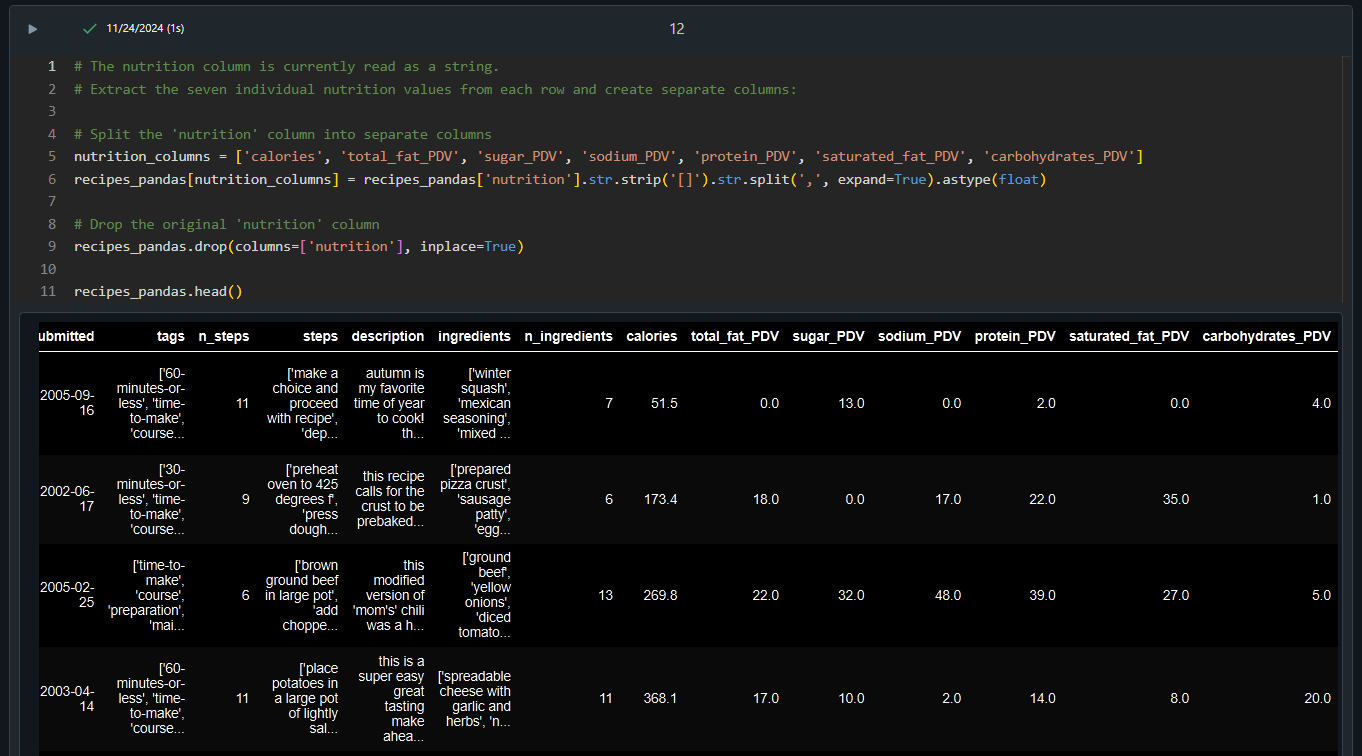
* **Task 2: Extract Nutrition Features**

Extracting the “Nutrition” Column” into individual features, specifically into 7 separate columns:

* calories
* total\_fat\_PDV
* sugar\_PDV
* sodium\_PDV
* protein\_PDV
* saturated\_fat\_PDV
* carbohydrates\_PDV

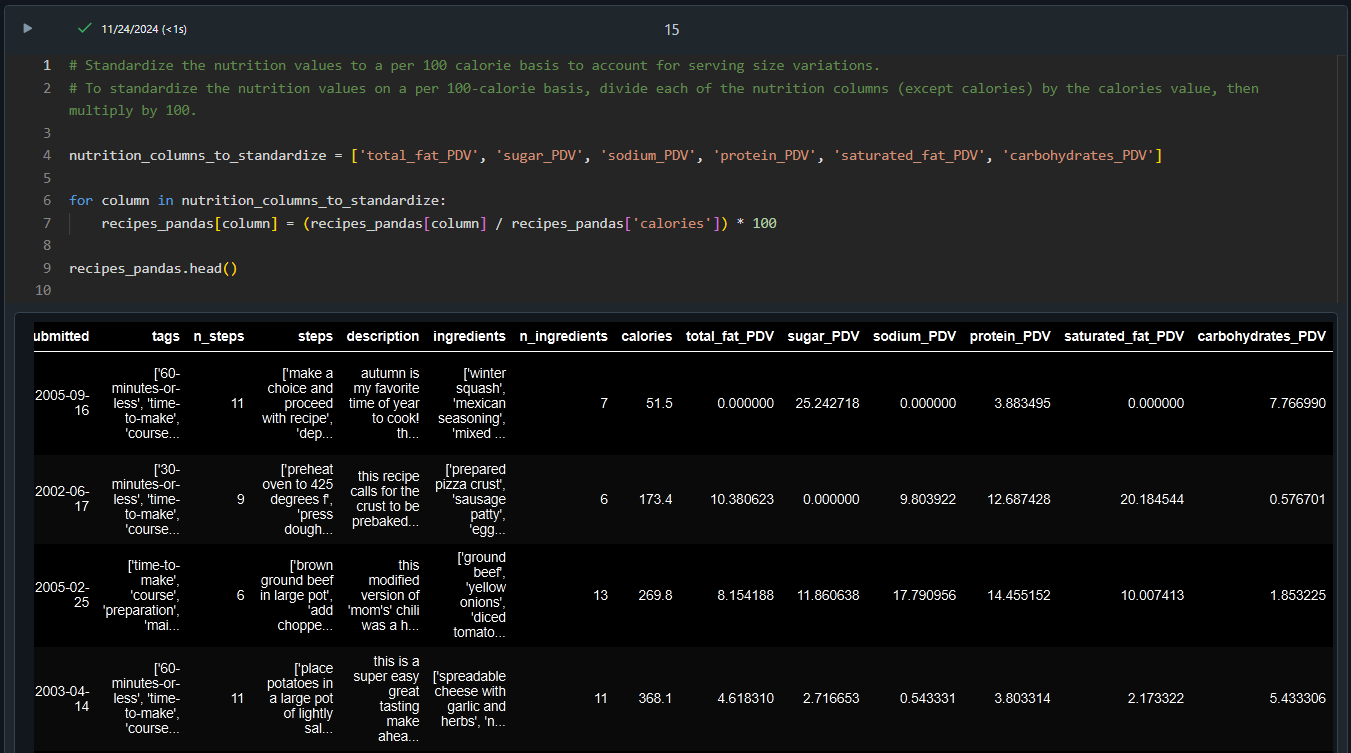
This helps us easily compare the dataset and usefully for the analysis processing.





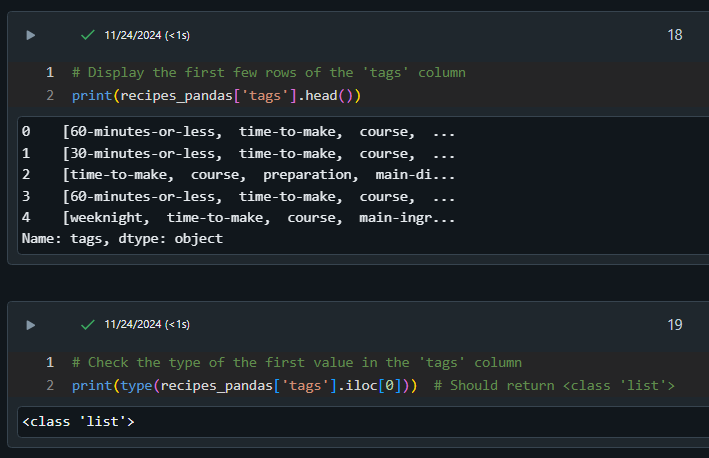
* **Task 3: Standardize Nutrition Values**

Standardize the nutrition values to a per 100 calories basis to account for serving size variations. Divide each of the nutrition columns (except calories) by the calories value, then multiply by 100. This helps avoid confusion between nutritional values ​​and allows for fair comparisons between recipes with different serving sizes.

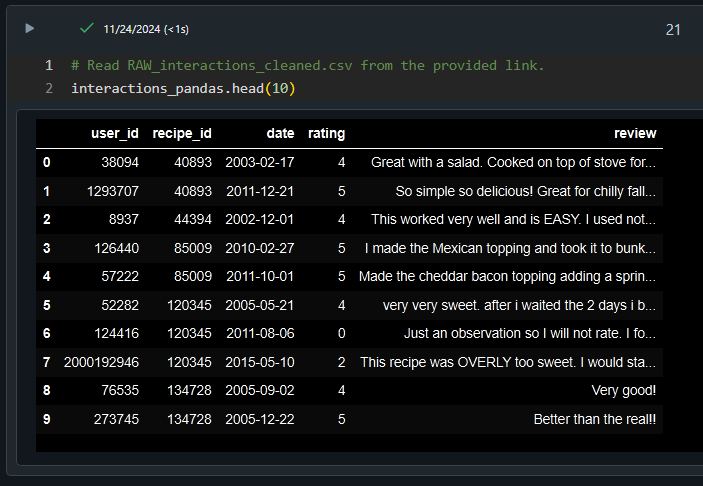


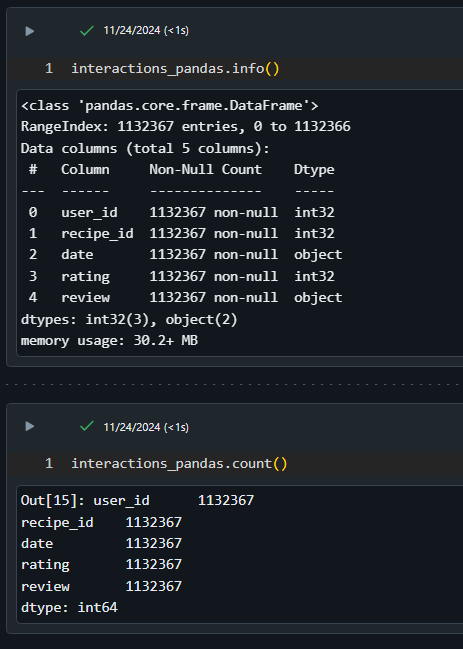
* **Task 4: Convert Tags to Array**

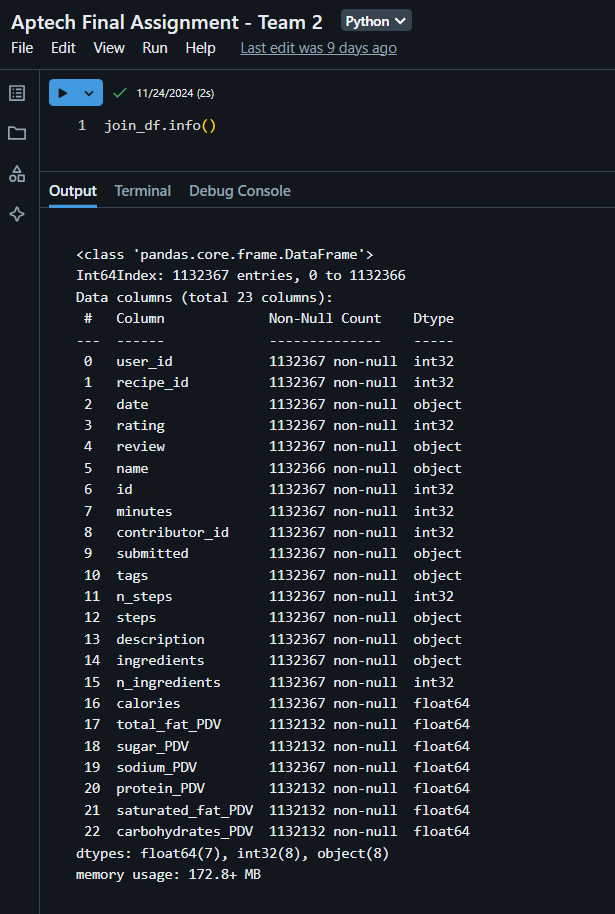
The tags column is currently read as a string. Convert it to an array of strings for ease of handling.



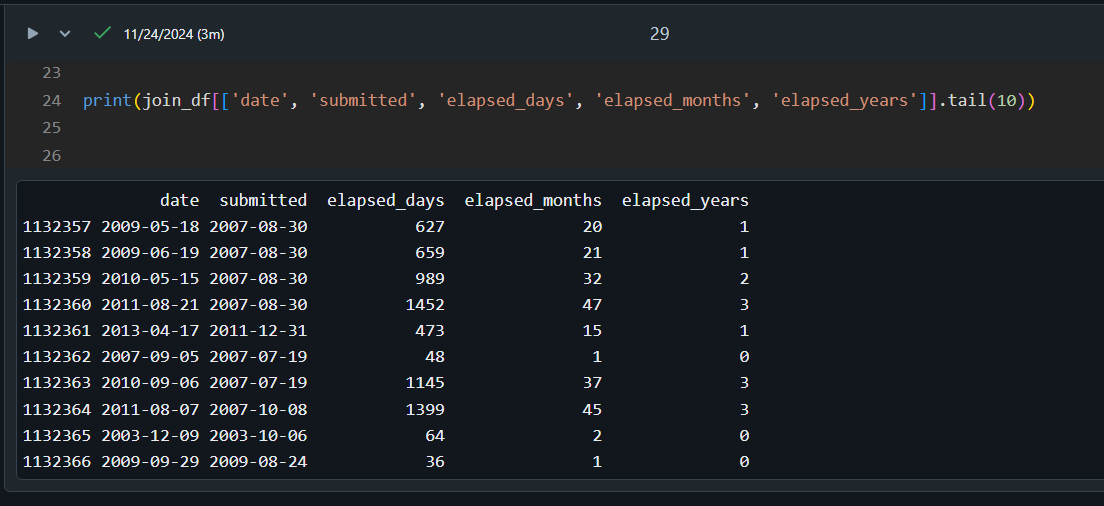
* **Task 5: Read and Join Interaction Data**







* **Task 6: Create Time-Based Features**

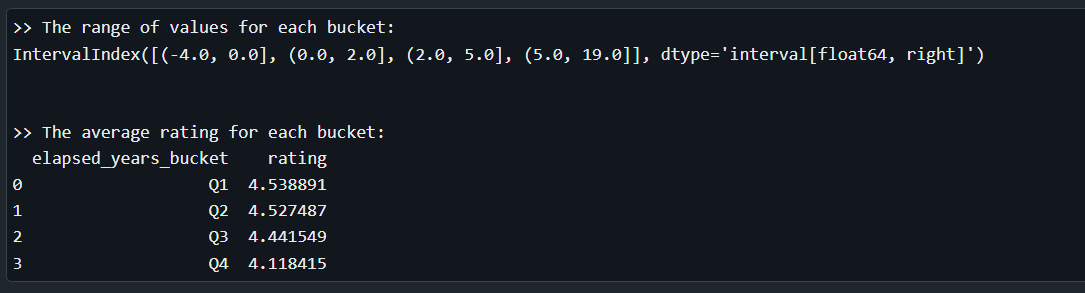


* **Task 7: Process Numerical Columns (Optional)**

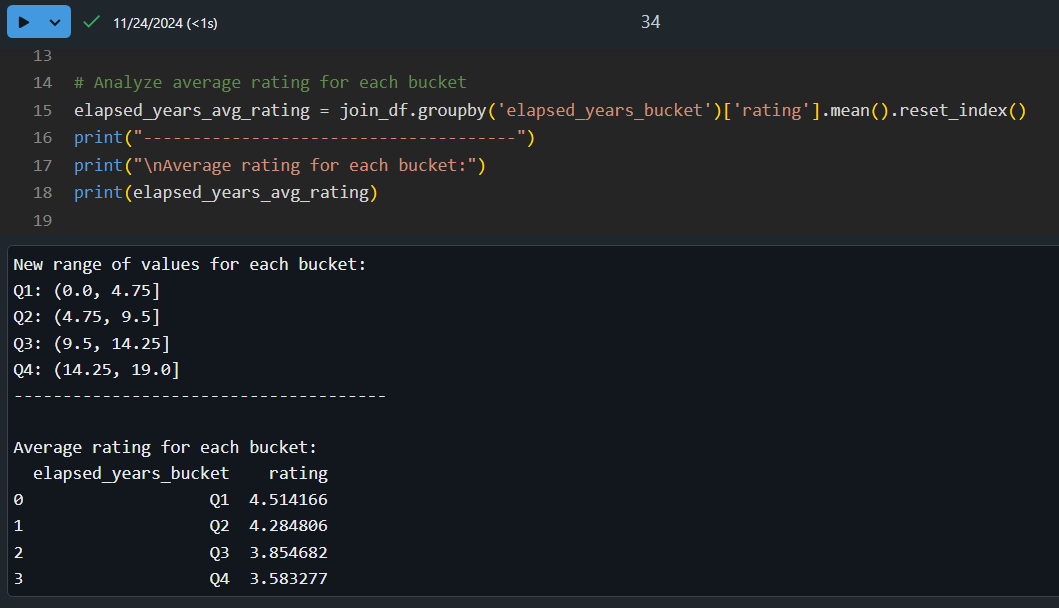
Introduce non-linearity to numerical columns by converting them into categorical columns using binning (percentile-based bucketing).

Analyze the average rating for each bucket to determine the usefulness of the bucketed column.

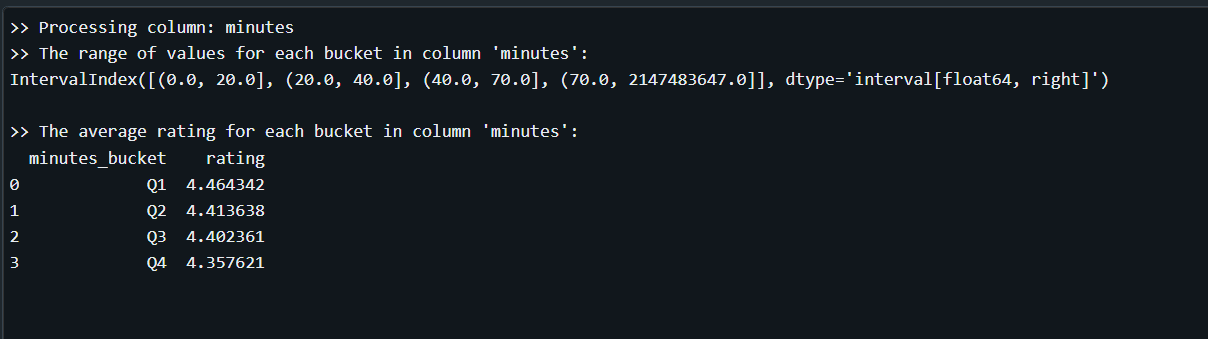
***1. Column years\_since\_submission\_on\_review\_date - elapsed\_years:***

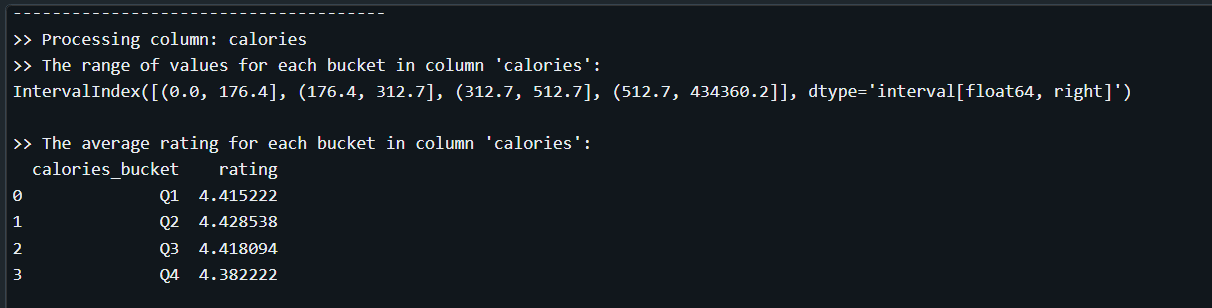


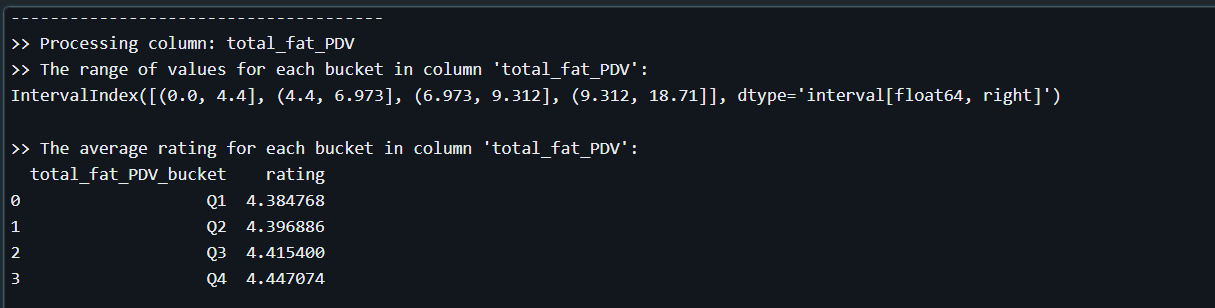
Next, drop rows with elapsed\_years < 0; and re-calculate the new range and average rating:

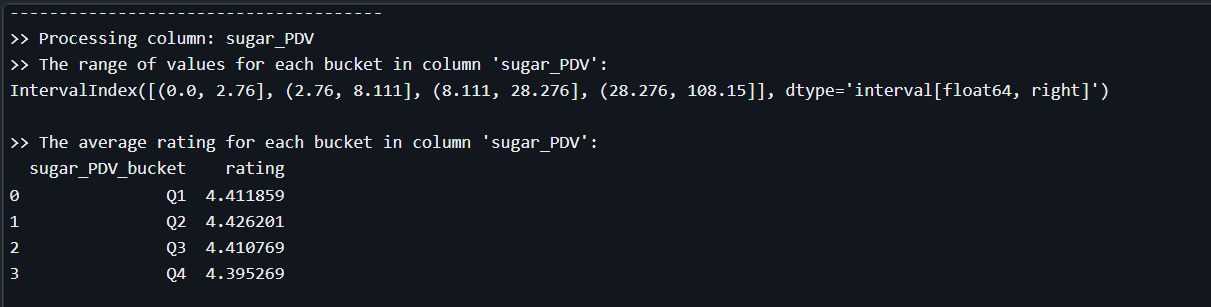


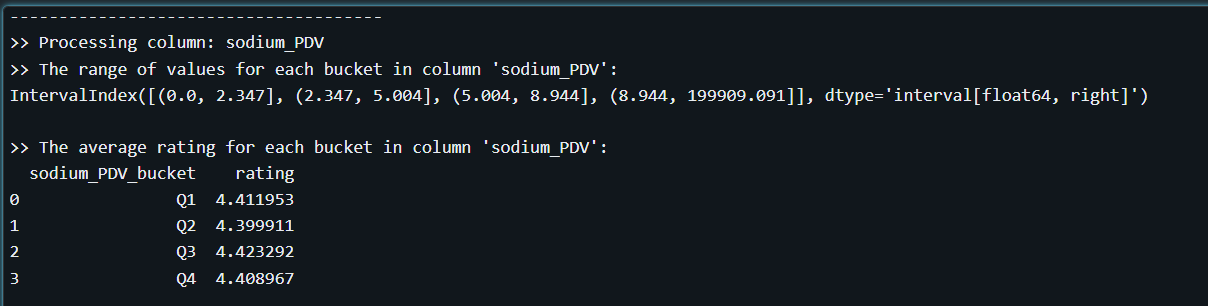
***2. For other columns:***

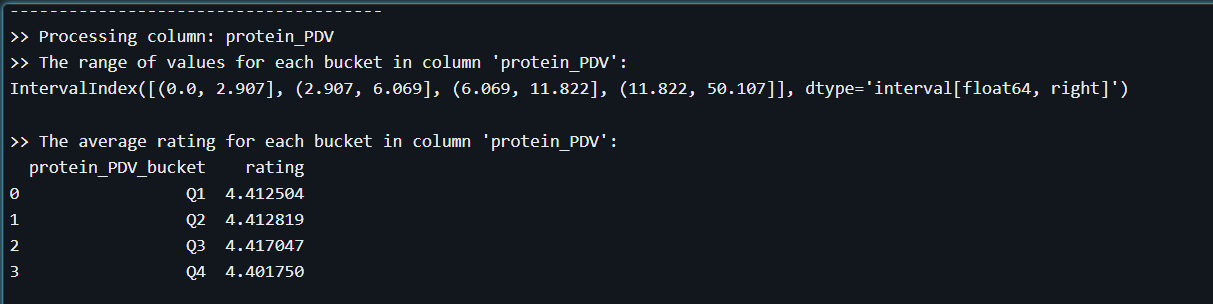


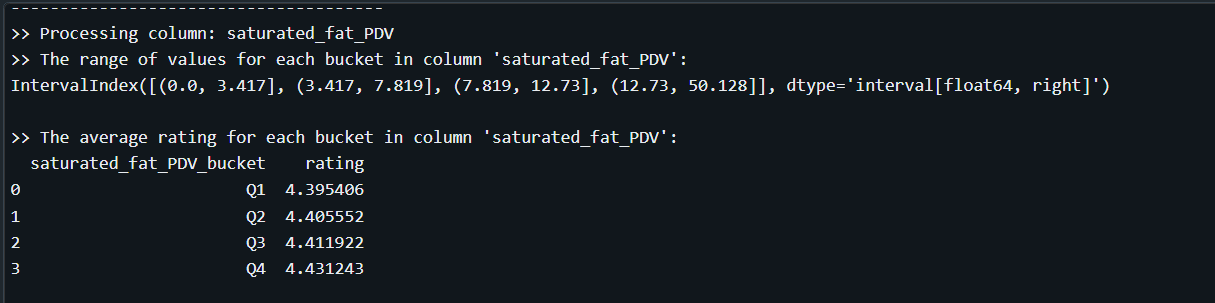


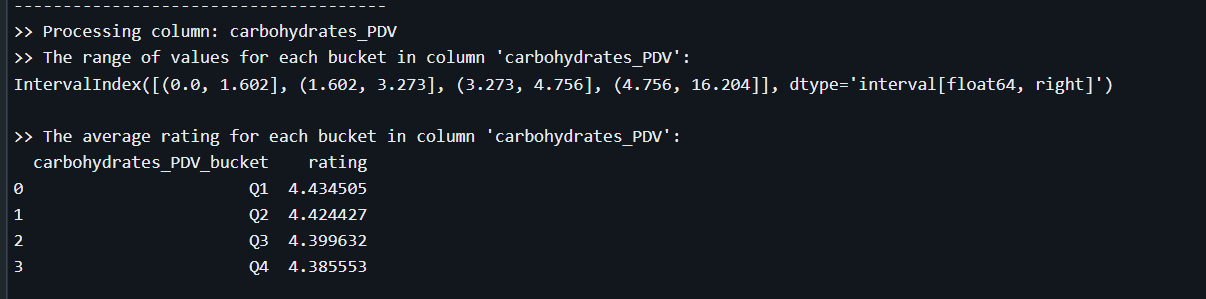




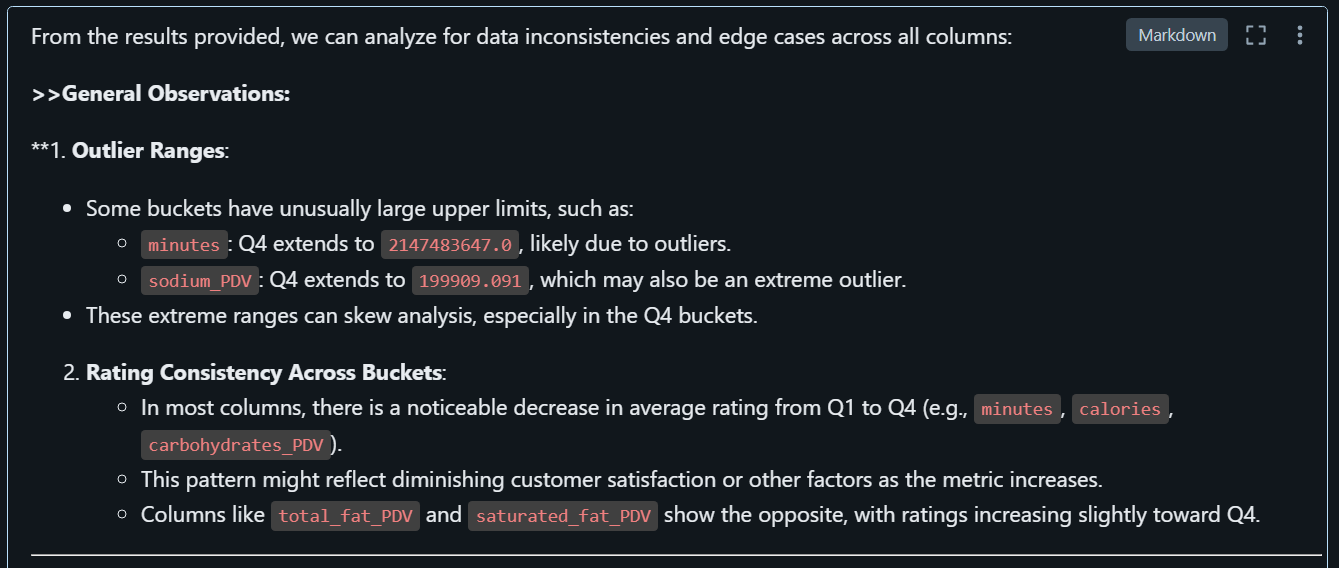


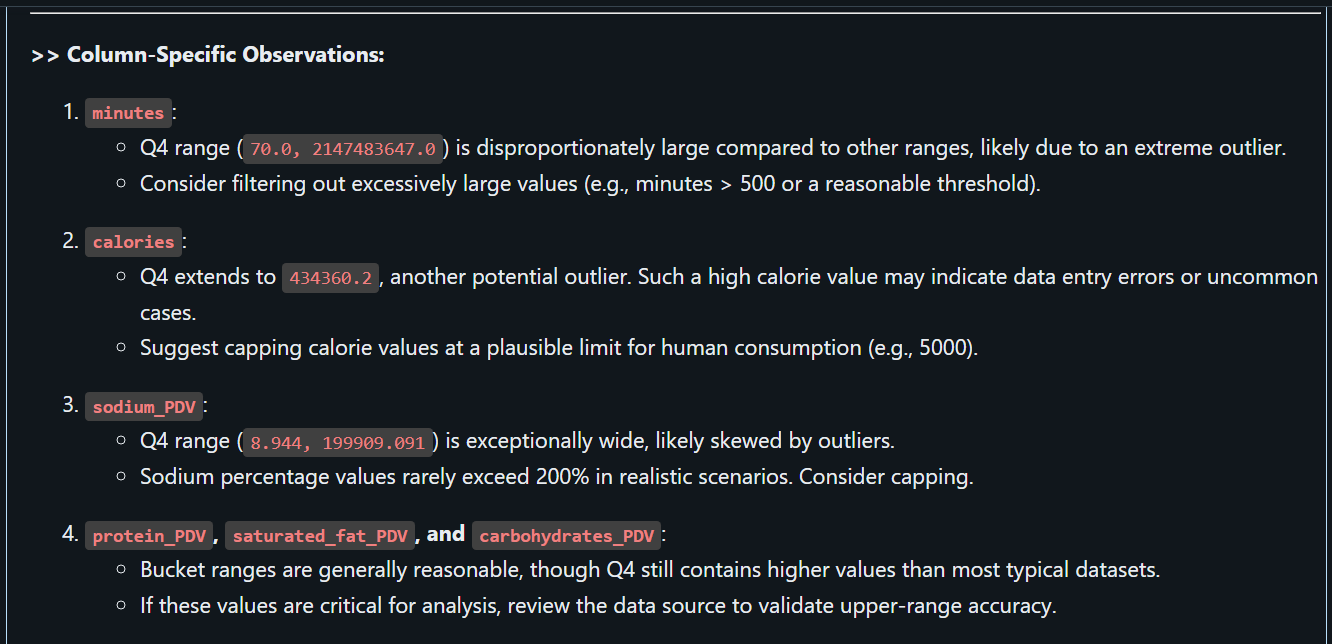


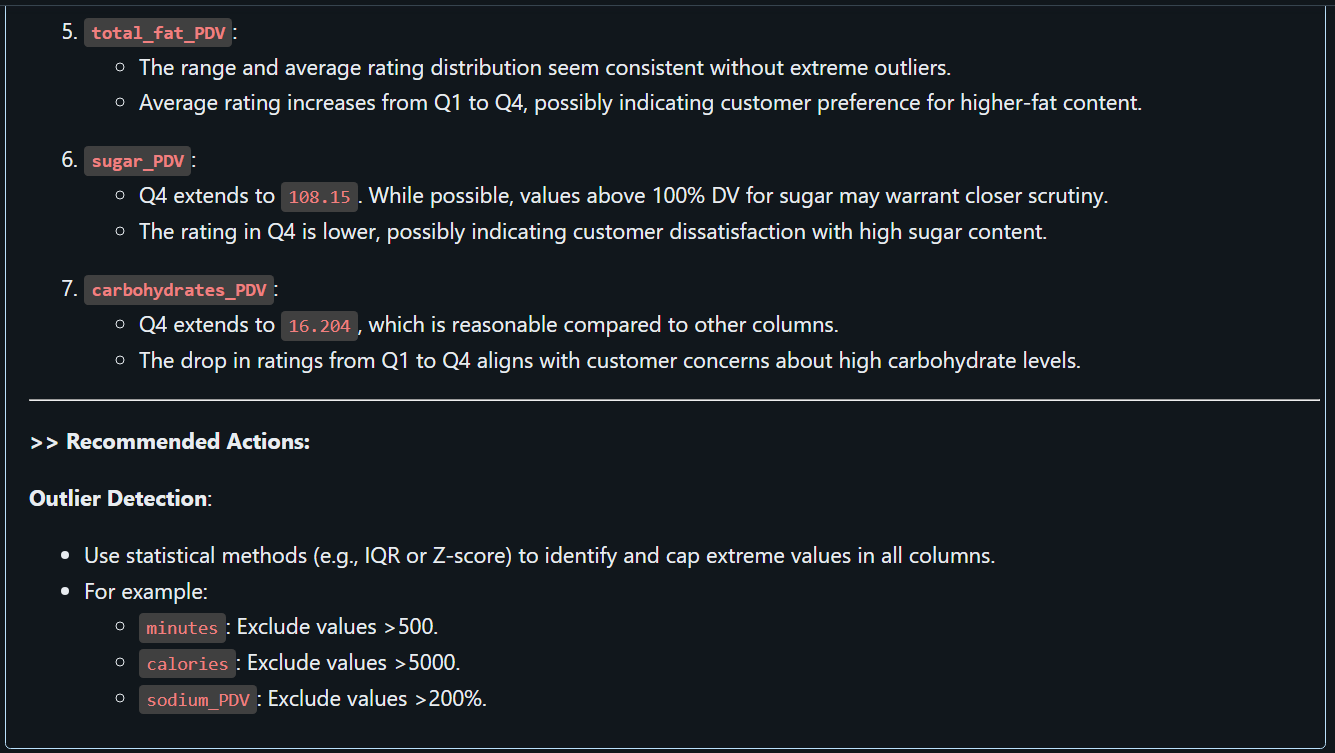




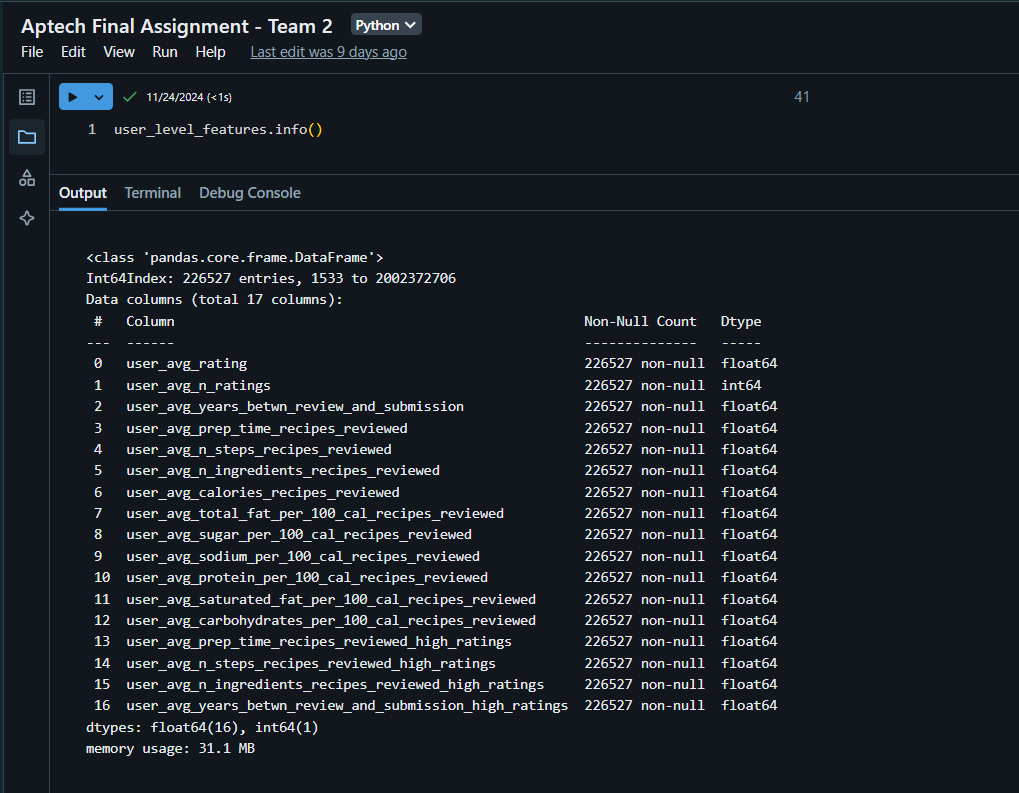


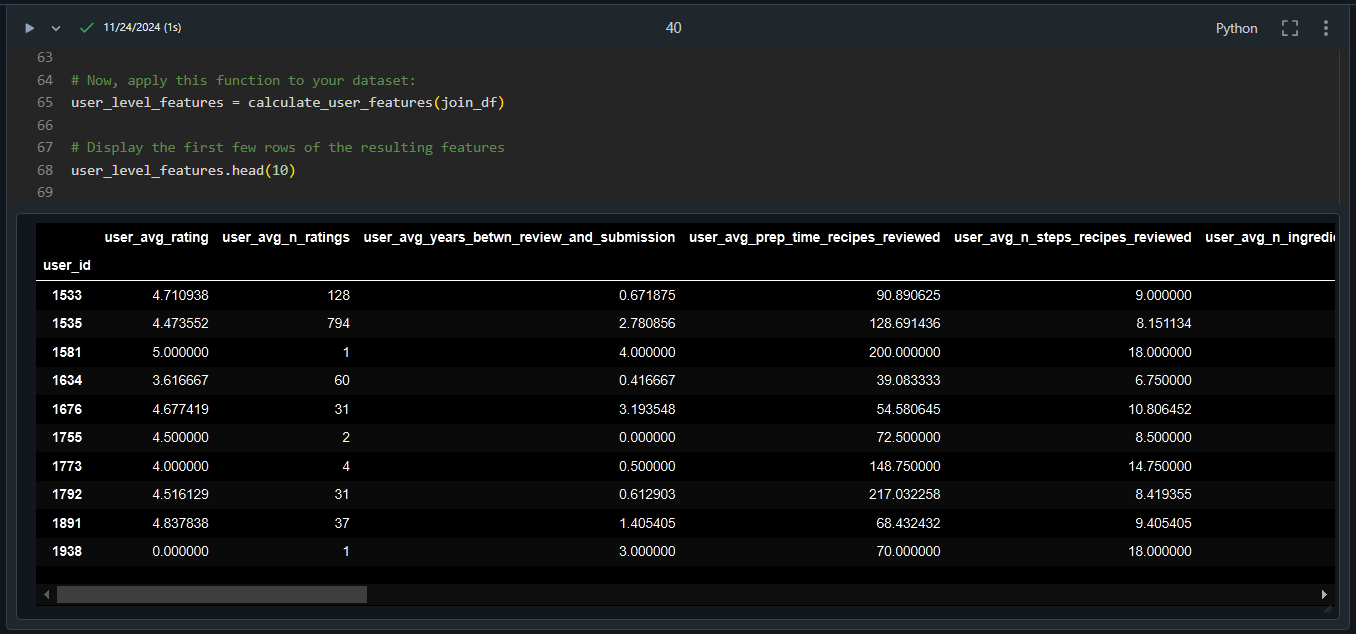


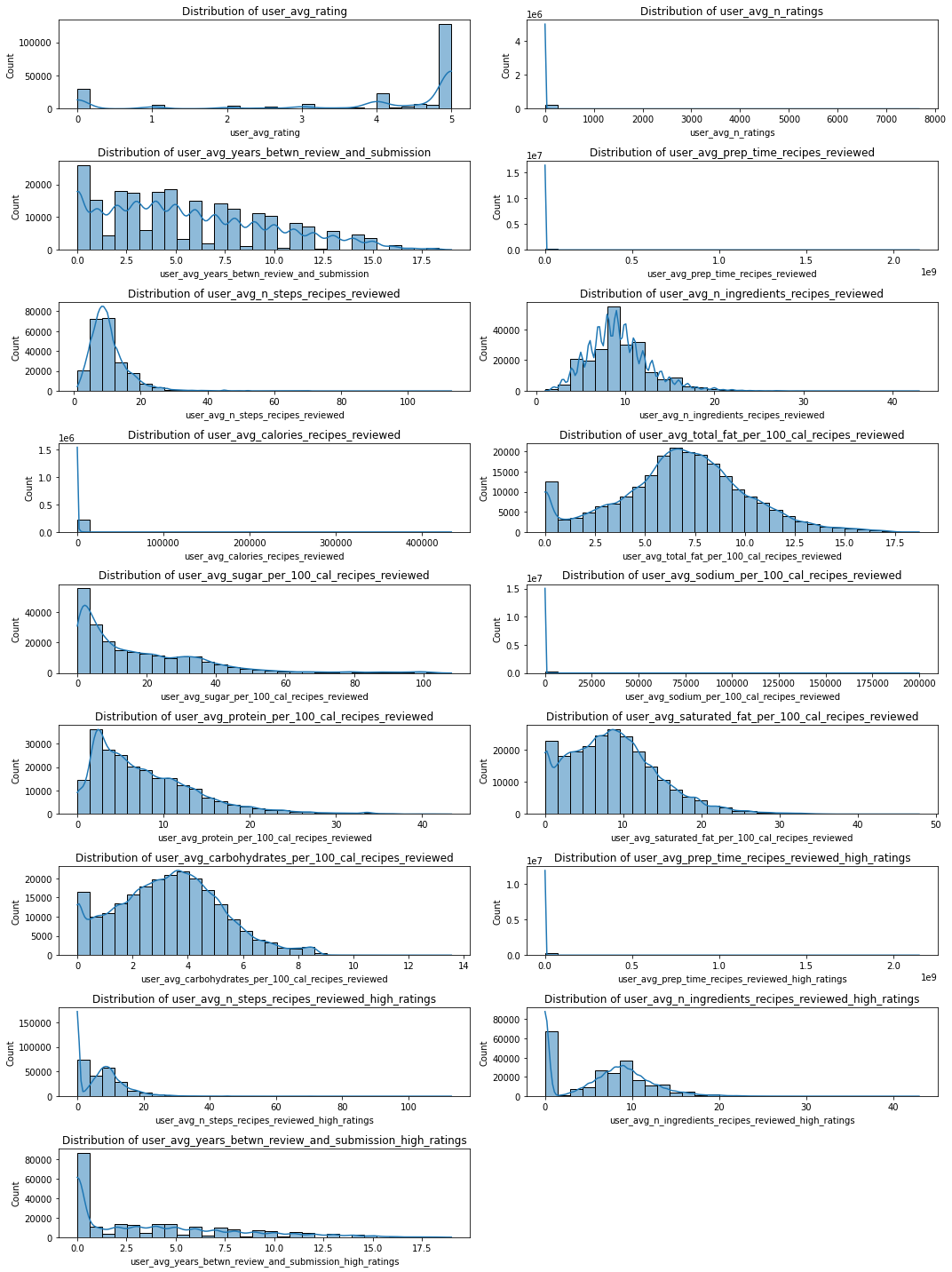




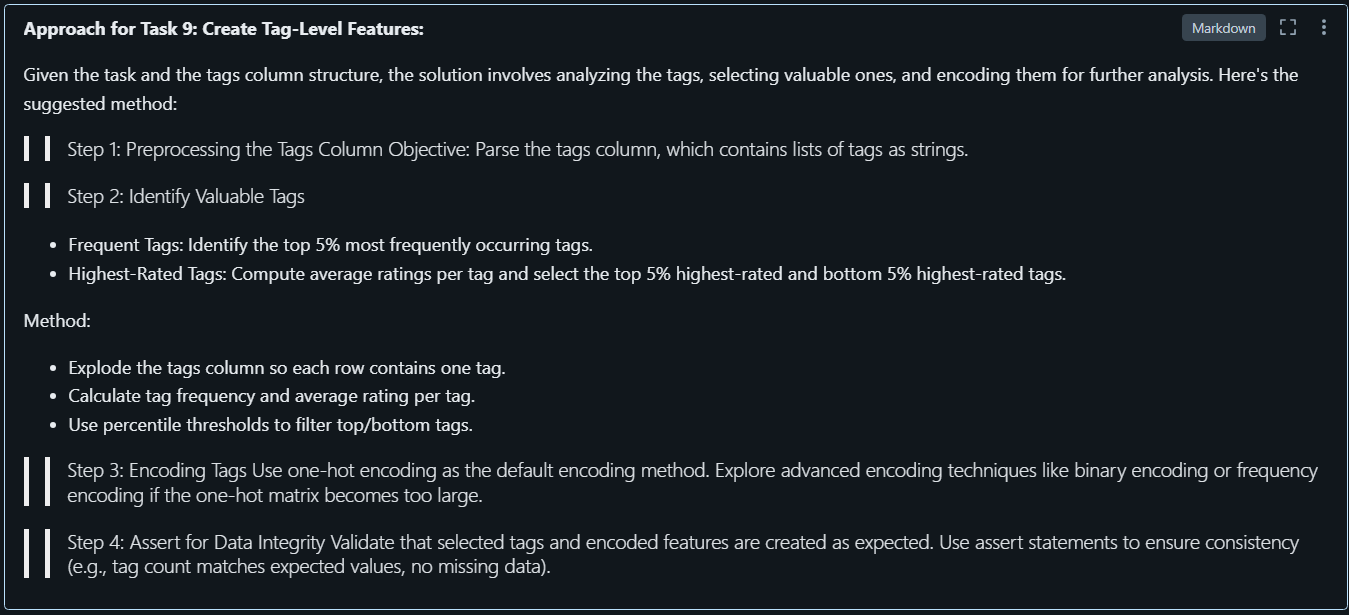
* **Task 8: Create User-Level Features (Optional)**

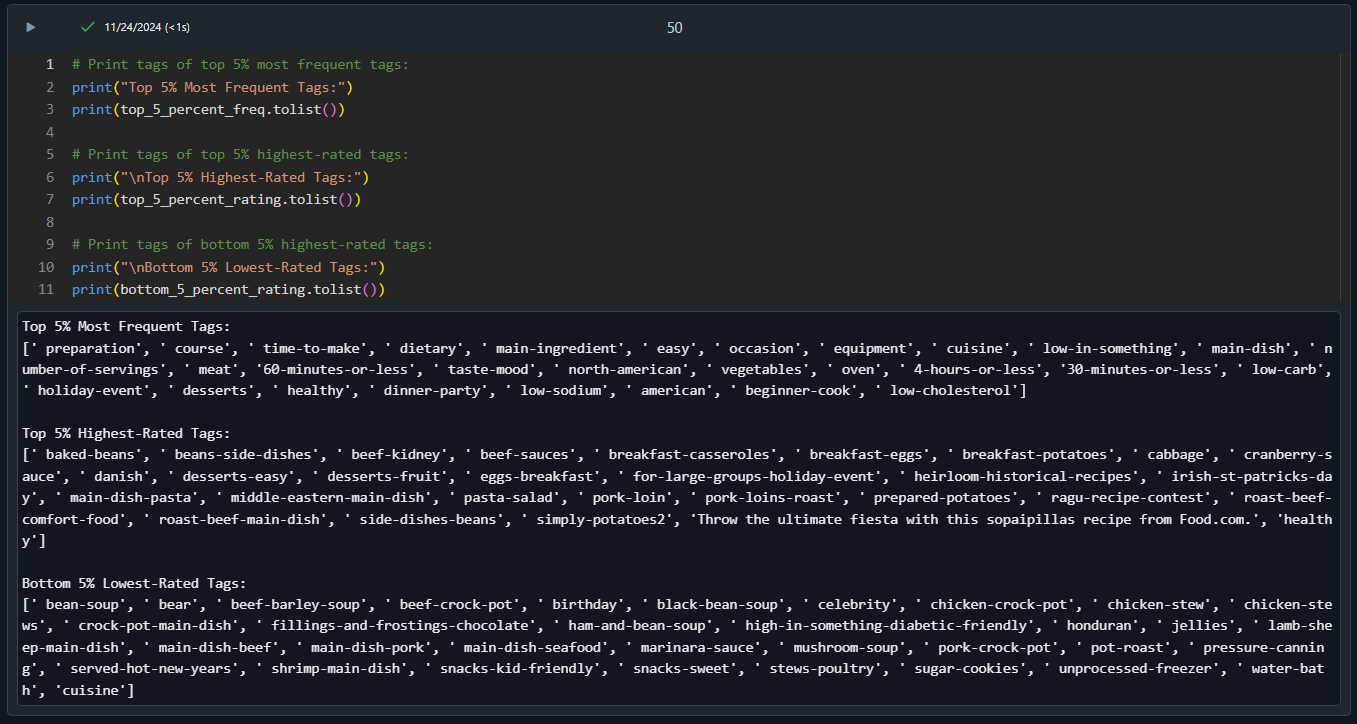


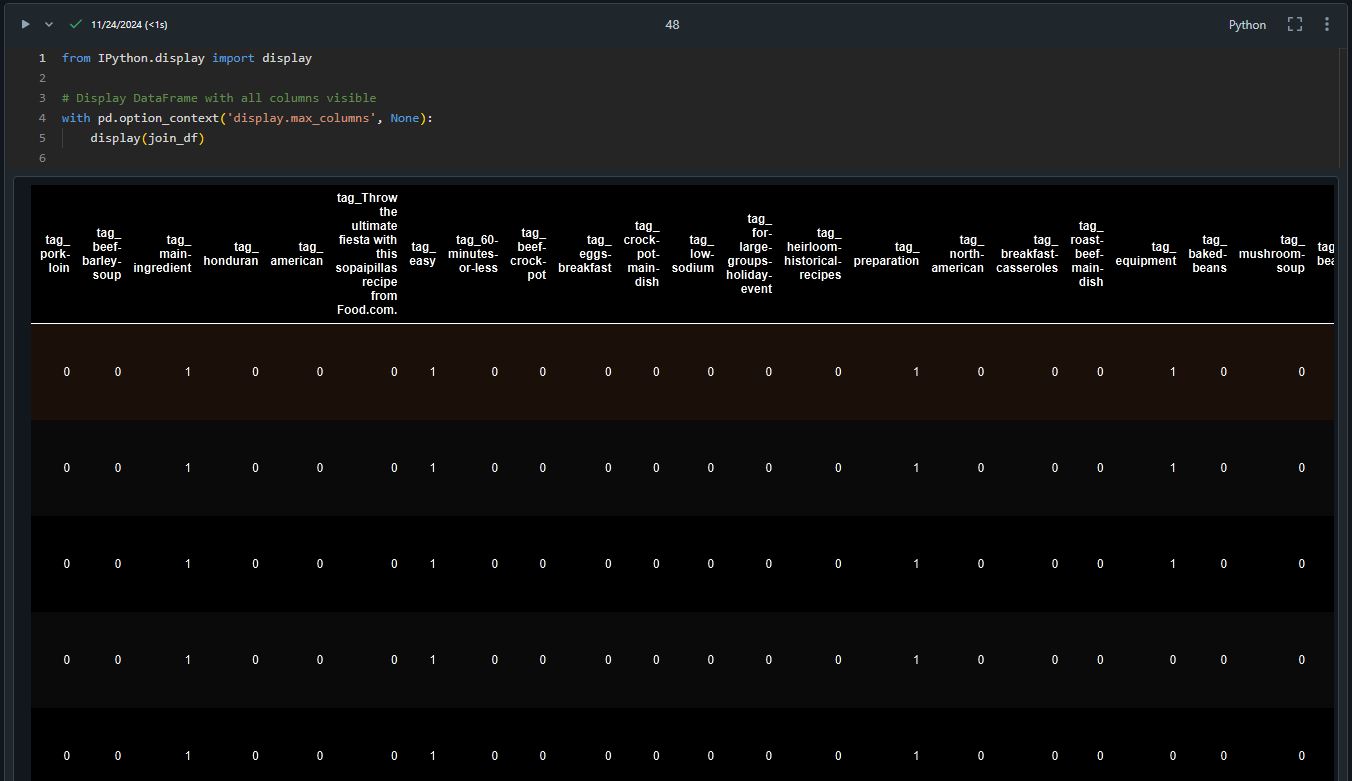




* **Task 9: Create Tag-Level Features (Optional)**







Here is the github link as requested: [LINK ADDED HERE]

Here is the data download link as requested: <https://raw-recipes-clean-upgrad.s3.amazonaws.com/RAW_recipes_cleaned.csv>

<https://raw-interactions-upgrad.s3.amazonaws.com/RAW_interactions_cleaned.csv>

**~~~ END ~~~**