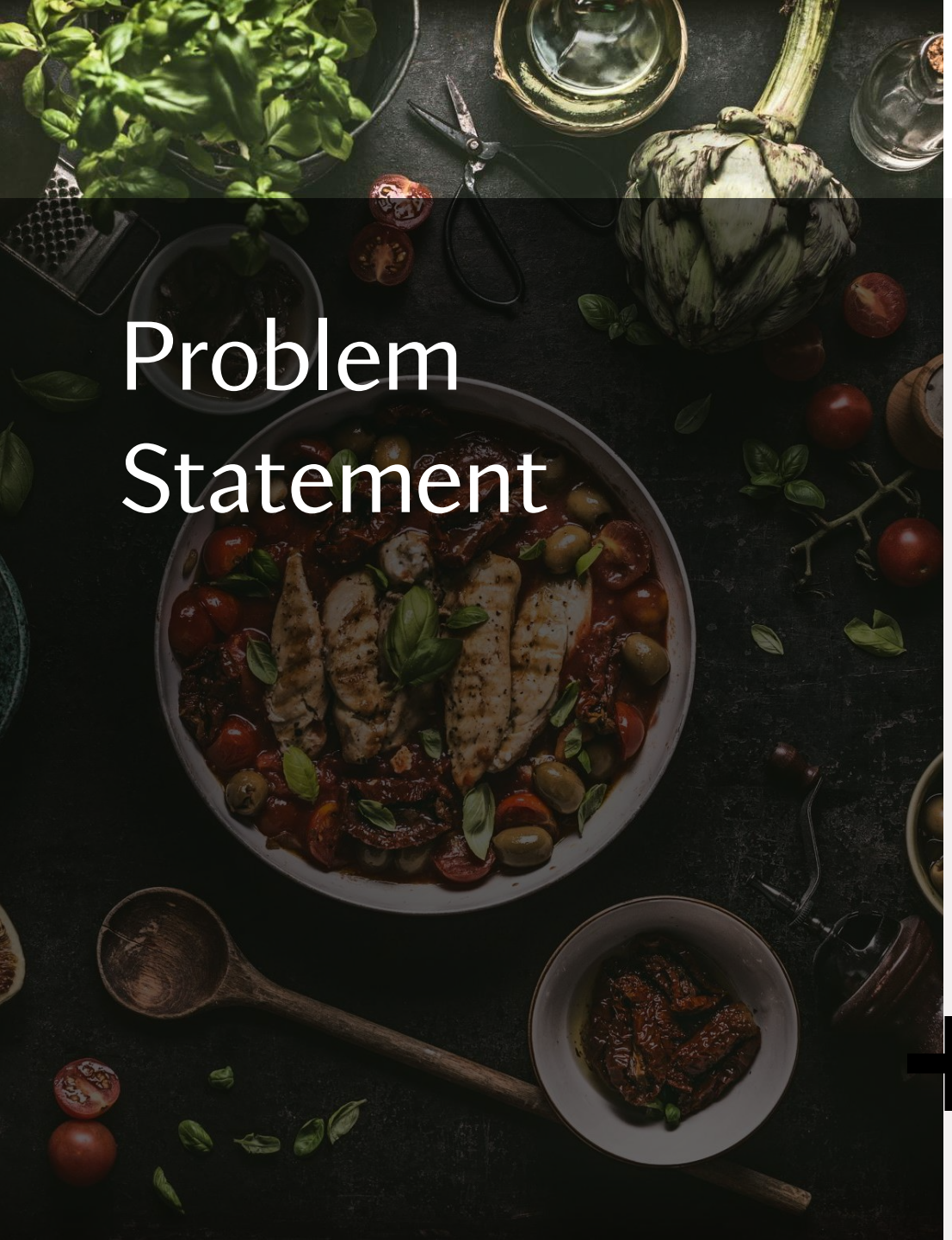
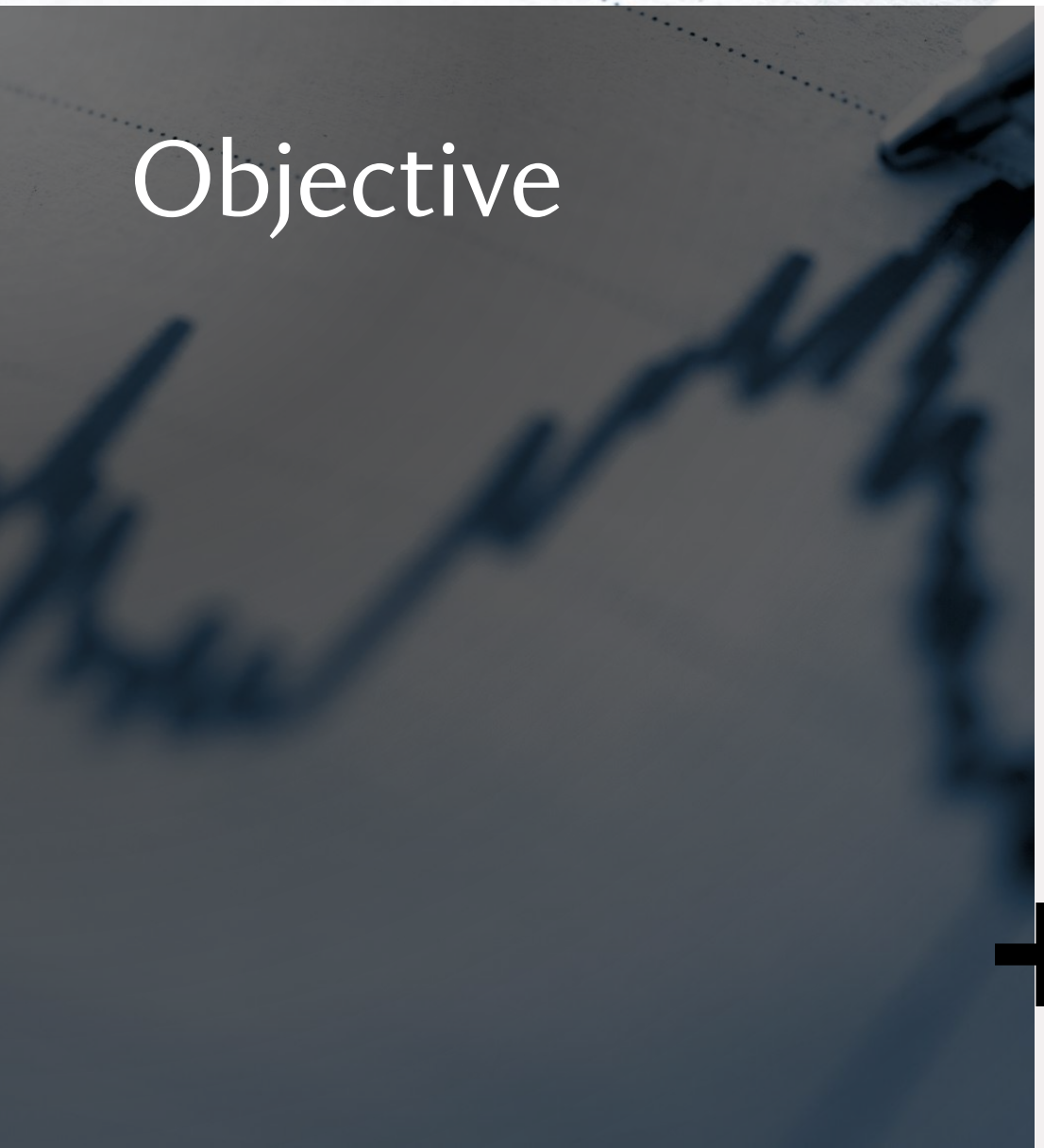
Recipe Recommender Assignment EDA Using PySpark

By Arnab Saha



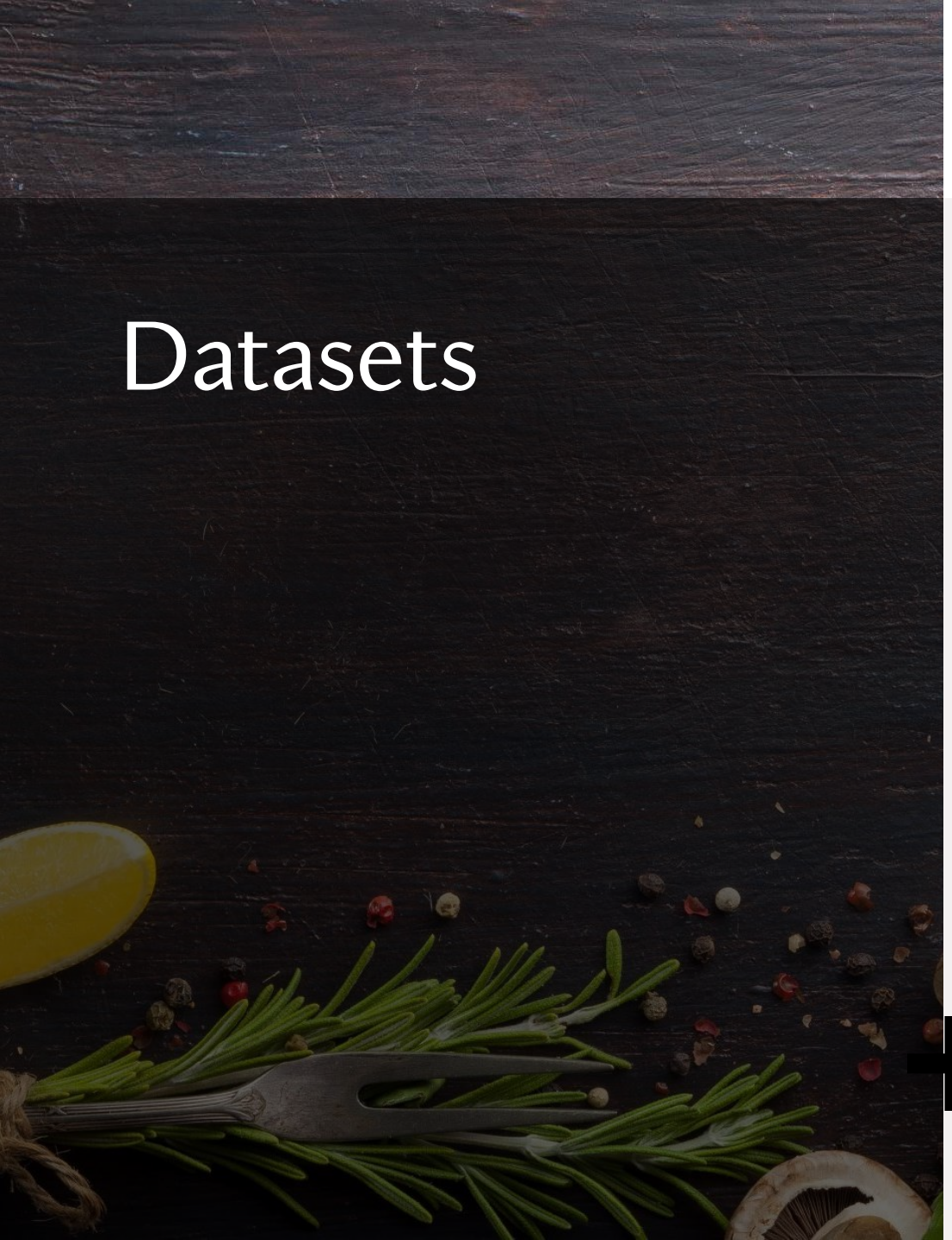
Problem Statement

- Food.com is a food recipe website where a user posts the food recipes and other users rate and comment on it.
- As 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.



Objective

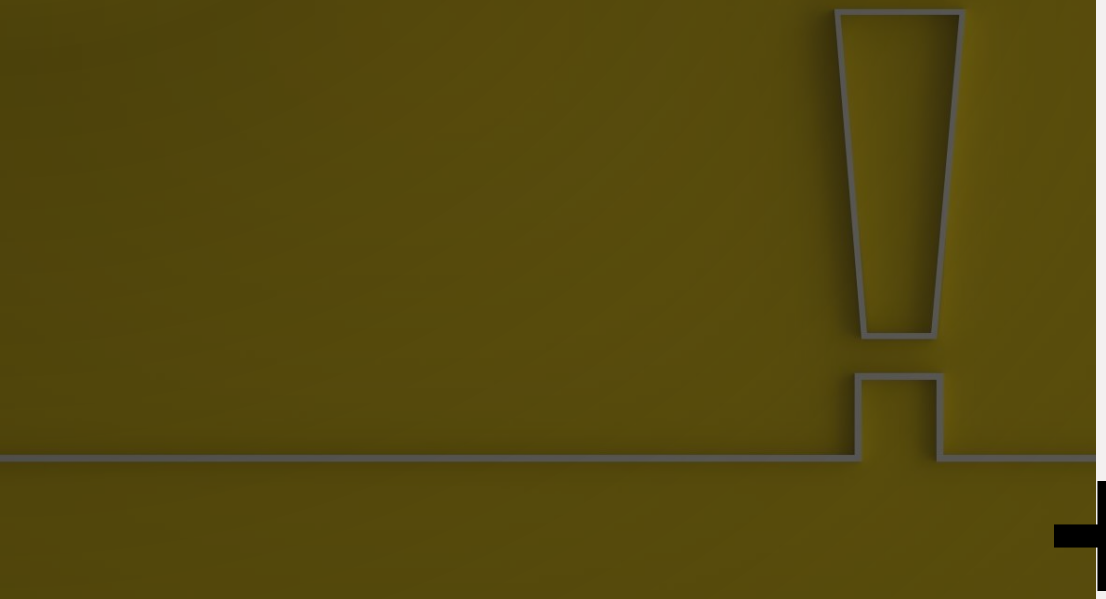
- Designing a recommender from scratch is a time-consuming task.
- In this assignment, we have used PySpark for the Data Analysis Purpose.
- Here we explored the raw data and created several features that can be used to build the recommender engine.



Datasets

- **Raw Recipe Data:**
  - https://raw-recipes-clean-upgrad.s3.amazonaws.com/RAW_recipes_cleaned.csv
- **Raw Ratings Data:**
  - https://raw-interactions-upgrad.s3.amazonaws.com/RAW_interactions_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.
- 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.

Steps



Part 1: Recipe Based Feature Extraction & EDA

- Task 01: Reading the first data file.
- Task 02: Extract individual features from the nutrition column.
- Task 03: Standardize the nutrition values.
- Task 04: Convert the tags column from a string to an array of strings.
- Task 05: Read the second data file.
- Task 06: Create time-based features.
- Task 07: Processing Numerical Columns and Visualising.

Part 2: User Based Feature Extraction & EDA

- Task 08: Create user-level features.
- Task 09: Create tag-level features and perform EDA.

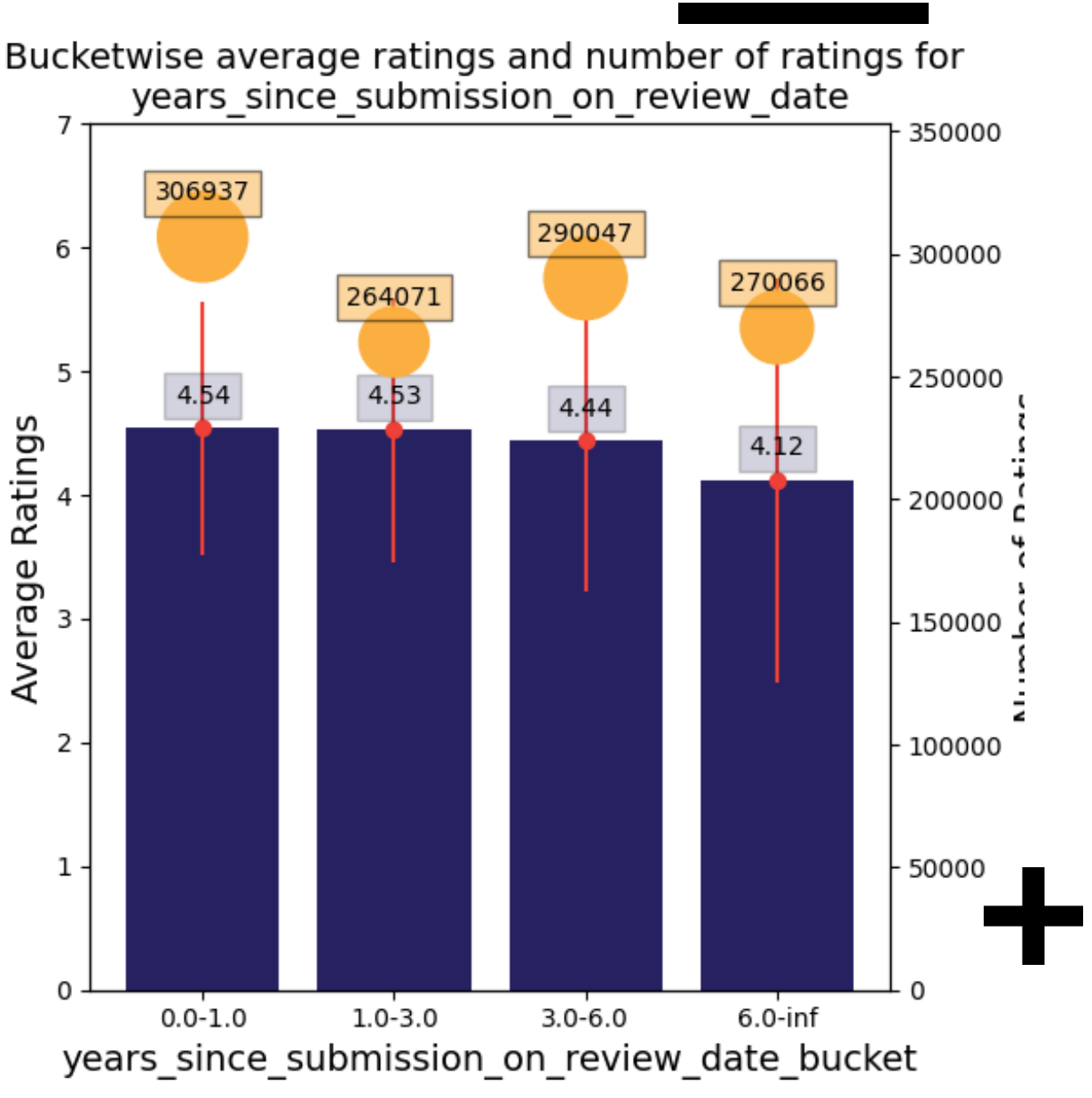
Part 3: Summary Analysis Report

- Summary Report.

Exploratory Data Analysis

Bucket-wise average ratings and number of ratings for "Years Since Submission on Review"

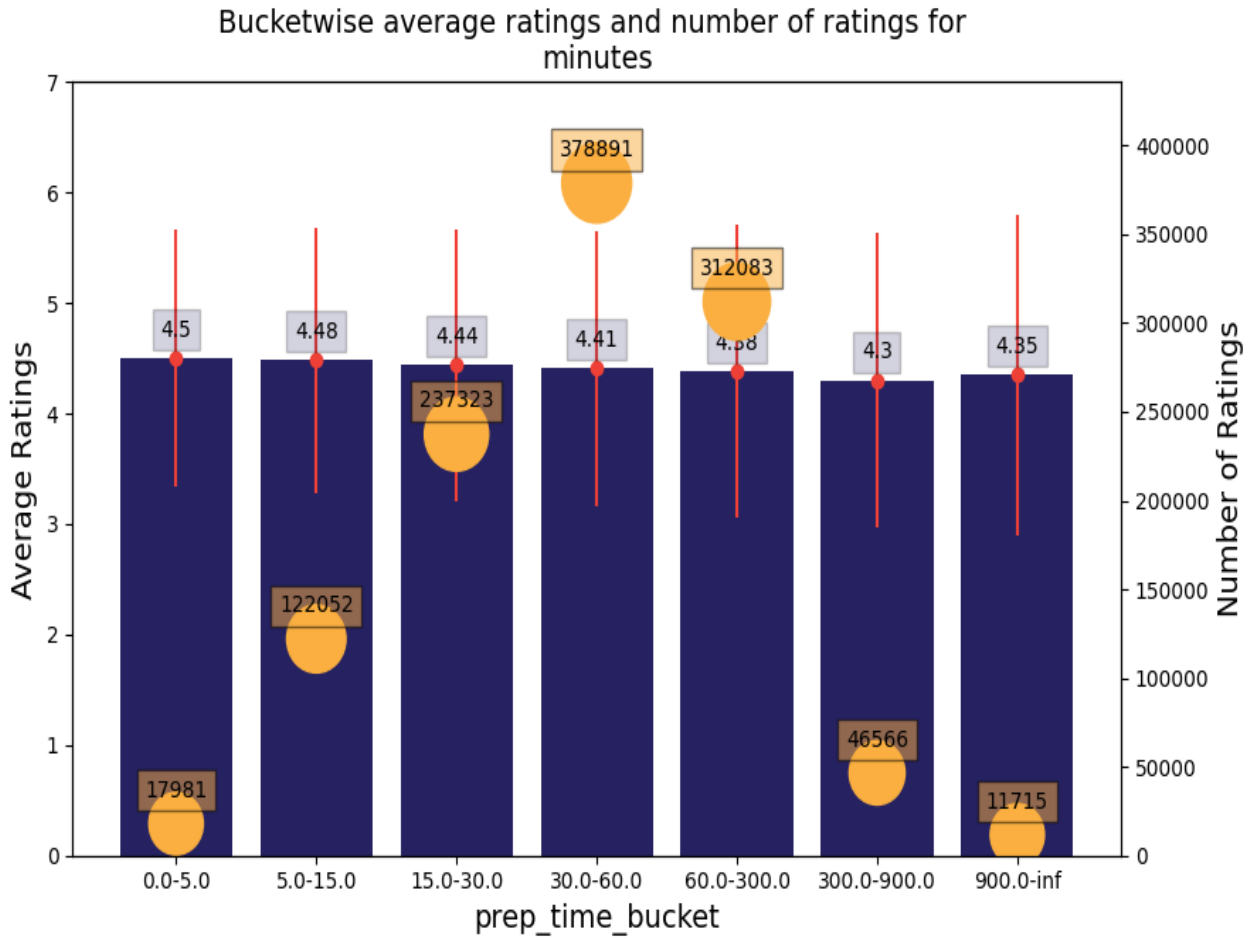
- It is the review time since submission of a recipe (in Years)
- It is found that Recipes more than 6 years old are rated low.



Exploratory Data Analysis

Bucket-wise average ratings and number of ratings for "Minutes"

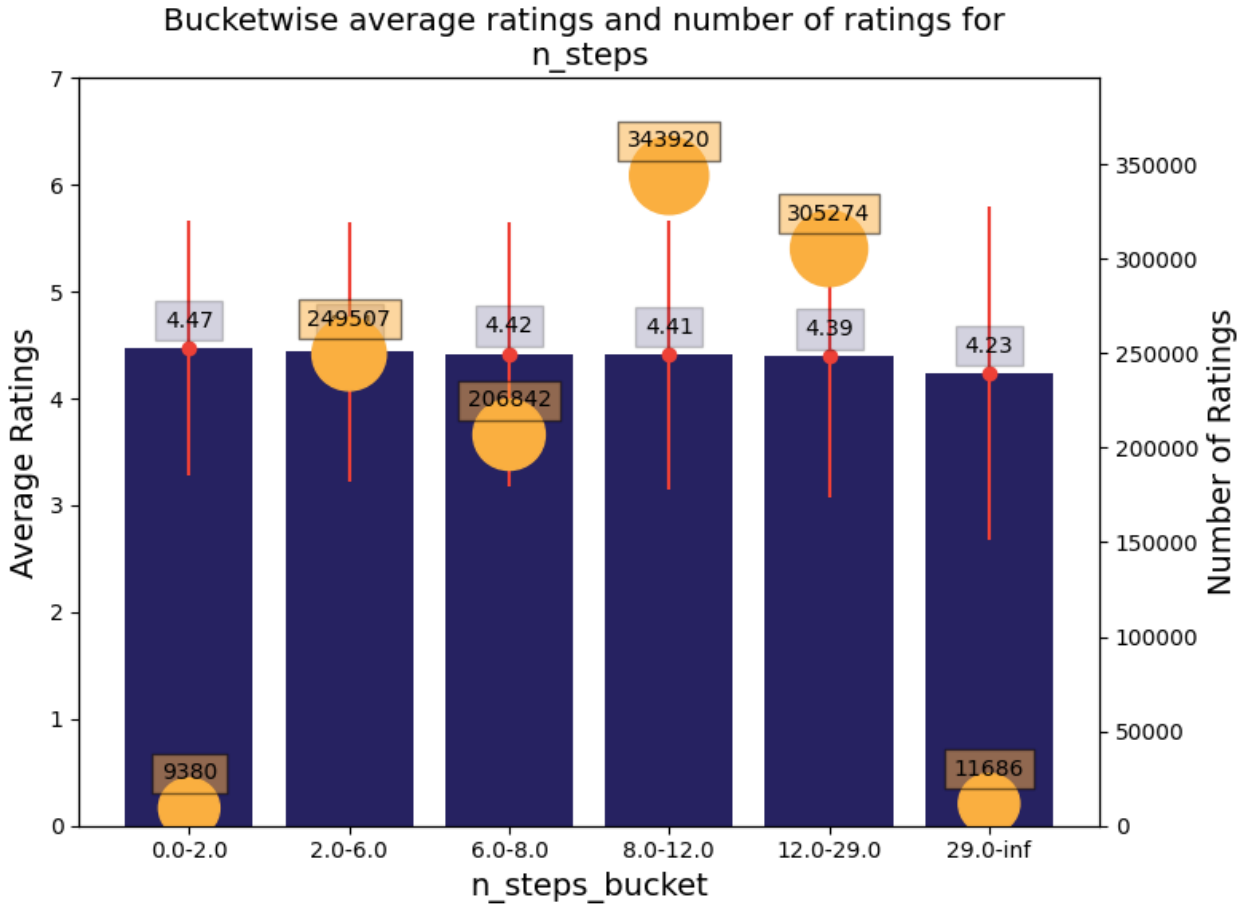
- It is the Time (in minutes) it takes to prepare recipe.
- It is found that Low preparation time is prefered.



Exploratory Data Analysis

Bucket-wise average ratings and number of ratings for "Number of Steps"

- It is the number of steps in the recipe.
- It is found that recipes with less than 2 steps are rated high.
- It is found that recipes with more than 29 steps are rated very low.



Exploratory Data Analysis

Based on number of user rated

Top n most rated Tags

```
added column: has_tag_easy
added column: has_tag_occasion
added column: has_tag_equipment
added column: has_tag_cuisine
added column: has_tag_low-in-something
added column: has_tag_main-dish
added column: has_tag_60-minutes-or-less
added column: has_tag_number-of-servings
added column: has_tag_meat
added column: has_tag_taste-mood
added column: has_tag_north-american
added column: has_tag_30-minutes-or-less
added column: has_tag_vegetables
added column: has_tag_oven
added column: has_tag_4-hours-or-less
added column: has_tag_low-carb
added column: has_tag_holiday-event
added column: has_tag_desserts
added column: has_tag_healthy
added column: has_tag_dinner-party
added column: has_tag_15-minutes-or-less
added column: has_tag_low-sodium
added column: has_tag_american
added column: has_tag_beginner-cook
added column: has_tag_low-cholesterol
added column: has_tag_low-calorie
added column: has_tag_inexpensive
added column: has_tag_comfort-food
added column: has_tag_kid-friendly
```

Bottom n least rated Tags

```
+--------------------+
|      individual_tag|
+--------------------+
|         snacks-sweet|
|lamb-sheep-main-dish|
|      cranberry-sauce|
|roast-beef-main-dish|
|                cabbage|
+--------------------+
```

Exploratory Data Analysis

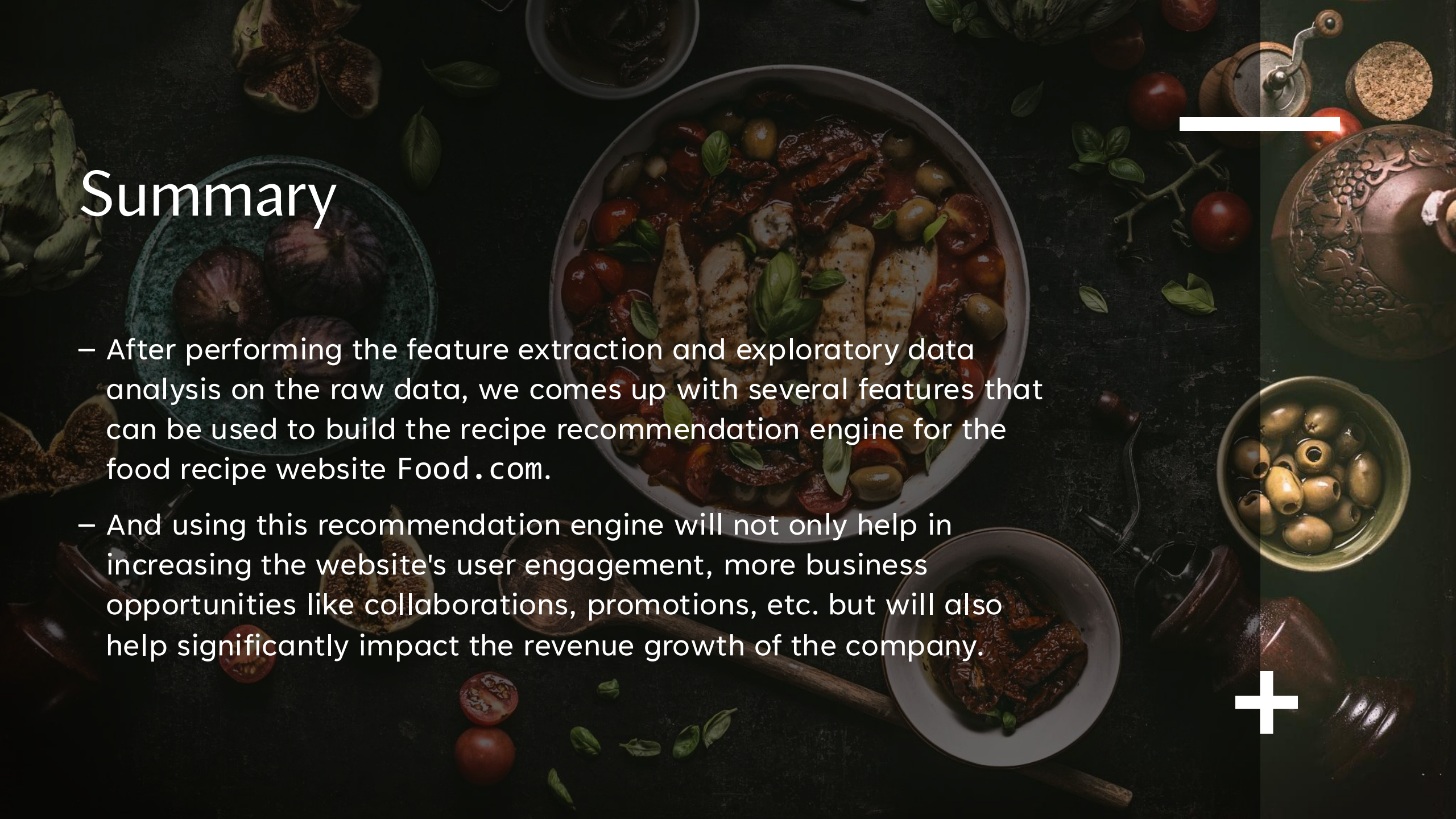
Based on average user rating

Top n rated Tags

```
added column: has_tag_ragu-recipe-contest
added column: has_tag_simply-potatoes2
added column: has_tag_non-alcoholic
added column: has_tag_a1-sauce
added column: has_tag_labor-day
added column: has_tag_punch
added column: has_tag_lettuces
added column: has_tag_cocktails
added column: has_tag_mashed-potatoes
added column: has_tag_smoothies
added column: has_tag_turkey-burgers
added column: has_tag_avocado
added column: has_tag_beverages
added column: has_tag_mango
added column: has_tag_asparagus
added column: has_tag_memorial-day
added column: has_tag_shakes
added column: has_tag_strawberries
added column: has_tag_omelets-and-frittatas
added column: has_tag_salsas
added column: has_tag_greek
added column: has_tag_salads
added column: has_tag_barbecue
added column: has_tag_australian
added column: has_tag_grilling
added column: has_tag_polynesian
```

Bottom n rated Tags

```
added column: has_tag_pressure-canning
added column: has_tag_honduran
added column: has_tag_unprocessed-freezer
added column: has_tag_birthday
added column: has_tag_jellies
added column: has_tag_water-bath
```



Summary

- After performing the feature extraction and exploratory data analysis on the raw data, we comes up with several features that can be used to build the recipe recommendation engine for the food recipe website Food.com.
- And using this recommendation engine will not only help in increasing the website's user engagement, more business opportunities like collaborations, promotions, etc. but will also help significantly impact the revenue growth of the company.

Thank You