



Recipe Recommender Assignment EDA

BY: MUKESH CHAUDHARY

NARGIS SHEIKH

NAVEEN GOTUR

Recipe Recommender Assignment

EDA

- ❖ Design a recommender system to recommend recipes to users based on their choice and the current recipe they are looking at for food.com by Extracting features from data.
- ❖ 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 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.

BUSINESS OBJECTIVE

- ❖ This recipe is to analyze user interactions with recipes in order to understand user preferences and identify patterns that can be used to improve the recipe recommendations for users.
- ❖ This can be accomplished by analyzing factors such as review time since submission, preparation time, number of steps, and number of ingredients, and determining which factors are most strongly correlated with high ratings.
- ❖ The resulting insights can be used to inform the development of a more effective recipe recommendation algorithm, which can help to increase user engagement and satisfaction with the recipe website or application.
- ❖ Additionally, this analysis can be used to identify potential areas for improvement in the recipe content and presentation, such as reducing the number of steps or ingredients in a recipe. The overall goal is to improve user experience and increase customer retention and acquisition.

- `from pyspark.sql import SparkSession`
- `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`

- ❖ This code imports various functions and types from the PySpark library, which is used for working with data in the Apache Spark framework.
- ❖ We can use these types and functions in our code to manipulate and analyze data stored in Spark Data Frames.
- ❖ We have included some test cases given below and we have completed the tasks.

SOLUTION METHODOLOGY

Data cleaning and data manipulations:

- ❖ Solution to all Tasks
- ❖ Read the data
- ❖ List of nutrition columns
- ❖ Extract individual features from the nutrition column
- ❖ Use string operations to remove the square brackets from the nutrition column
- ❖ Split the nutrition column into seven individual columns and cast the new columns to float values
- ❖ Nutrition column split into multiple
- ❖ We have included some test cases given below. You can use them to check if you have completed the task correctly

DATA MANIPULATION

- ❖ Standardize the nutrition values
- ❖ By converting the nutrition values from absolute to relative terms, we ensure that portion size is not a factor in the analysis.
- ❖ All nutrition columns standardized to per 100 calories. We have included some test cases given below. You can use them to check if you have completed the task correctly
- ❖ Complete the code in the following cell
- ❖ Convert the tags column from a string to an array of strings
- ❖ Join Recipe Data to Review Data and Read the second data file
- ❖ Create time-based features
- ❖ Save the data we have created so far in a parquet file.

`('s3a://upgradfoodrecsysdir/interaction_level_df_processed.parquet')`

- ❖ Has a header row and that the data types for all columns should be inferred automatically. The file is located at “s3://mybuckethandson/RAW_recipes_cleaned.csv”

('s3://recipe-assignment/data/interaction_level_df_processed.parquet')

('s3://recipe-assignment/data/interaction_level_df_postEDA.parquet')

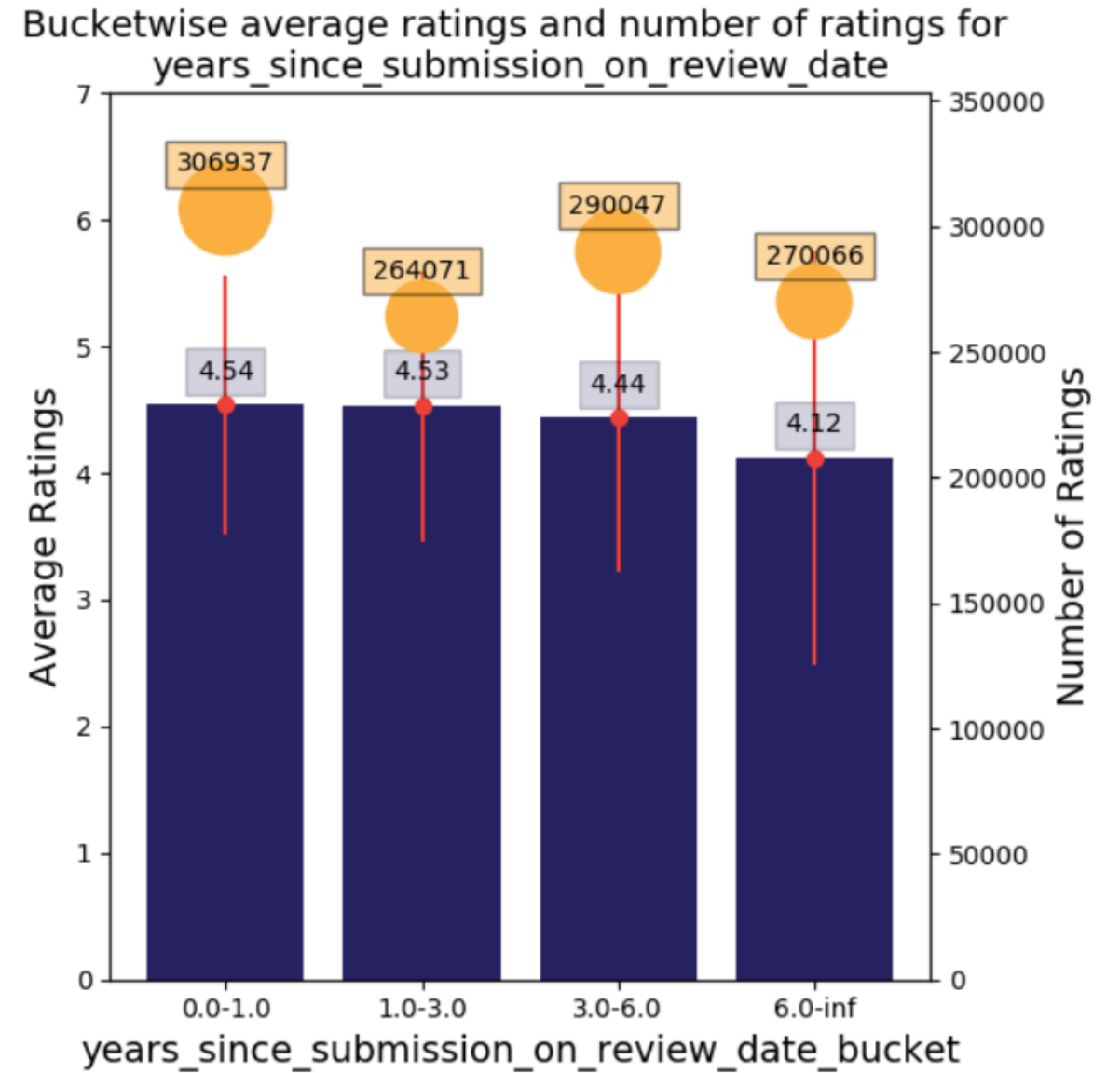
('s3://recipe-assignment/data/interaction_level_df_ModelReady.parquet')

- ❖ Such values and Null values are treated as Not Declared and used for further analysis.
- ❖ Numerical Missing values have been dropped.
- ❖ Outlier Treatment of TotalVisits and Page Views Per Visit.
- ❖ Observed that major part of null values in "Page Views Per Visit", "TotalVisits" are Converted. So, imputing median values of them to null values.

EDA

Years since submission on review date

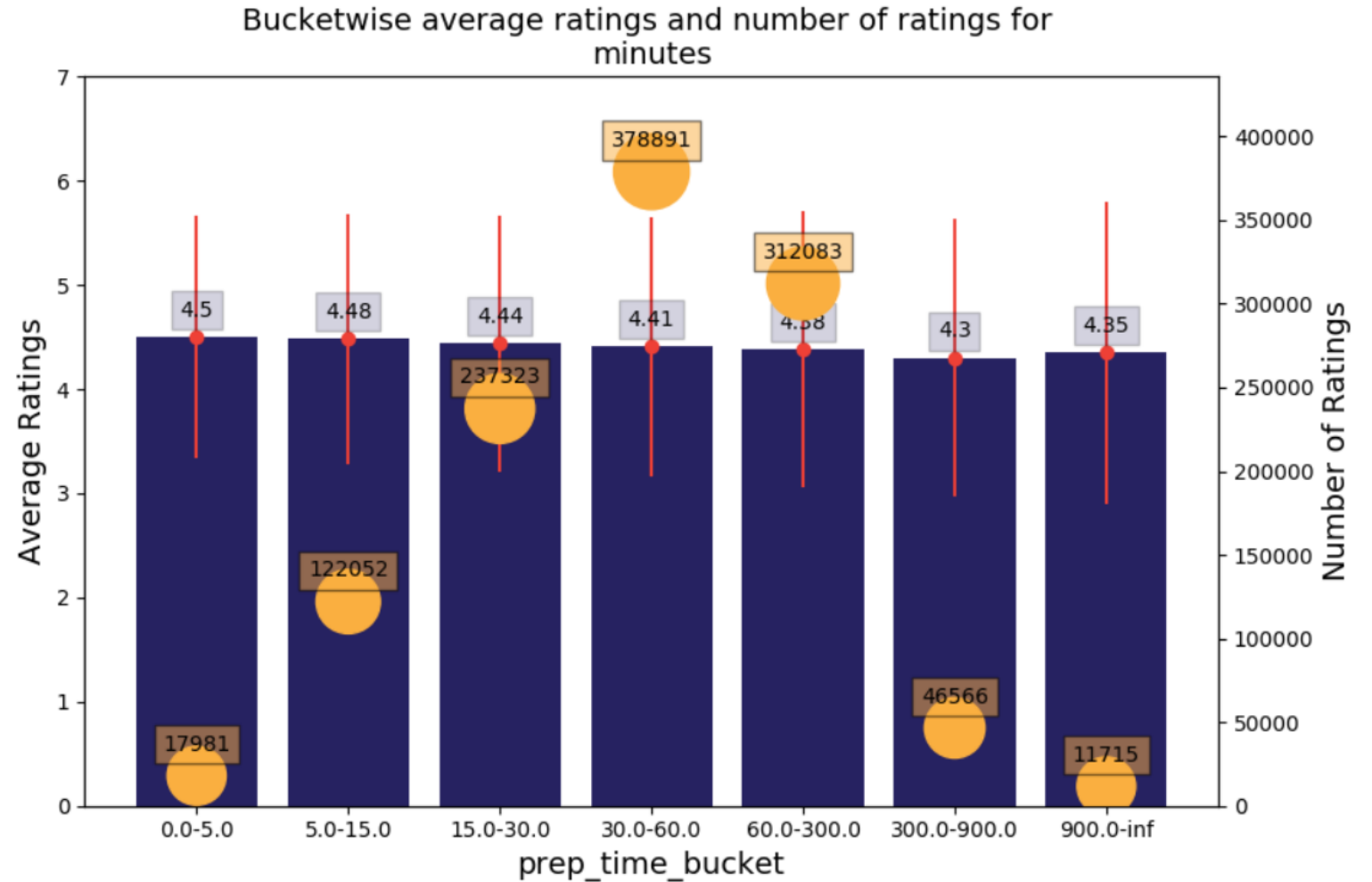
- ❖ [Review Time Since Submission]
- ❖ Recipes more than 6 years old are rated low



EDA

Preparation time

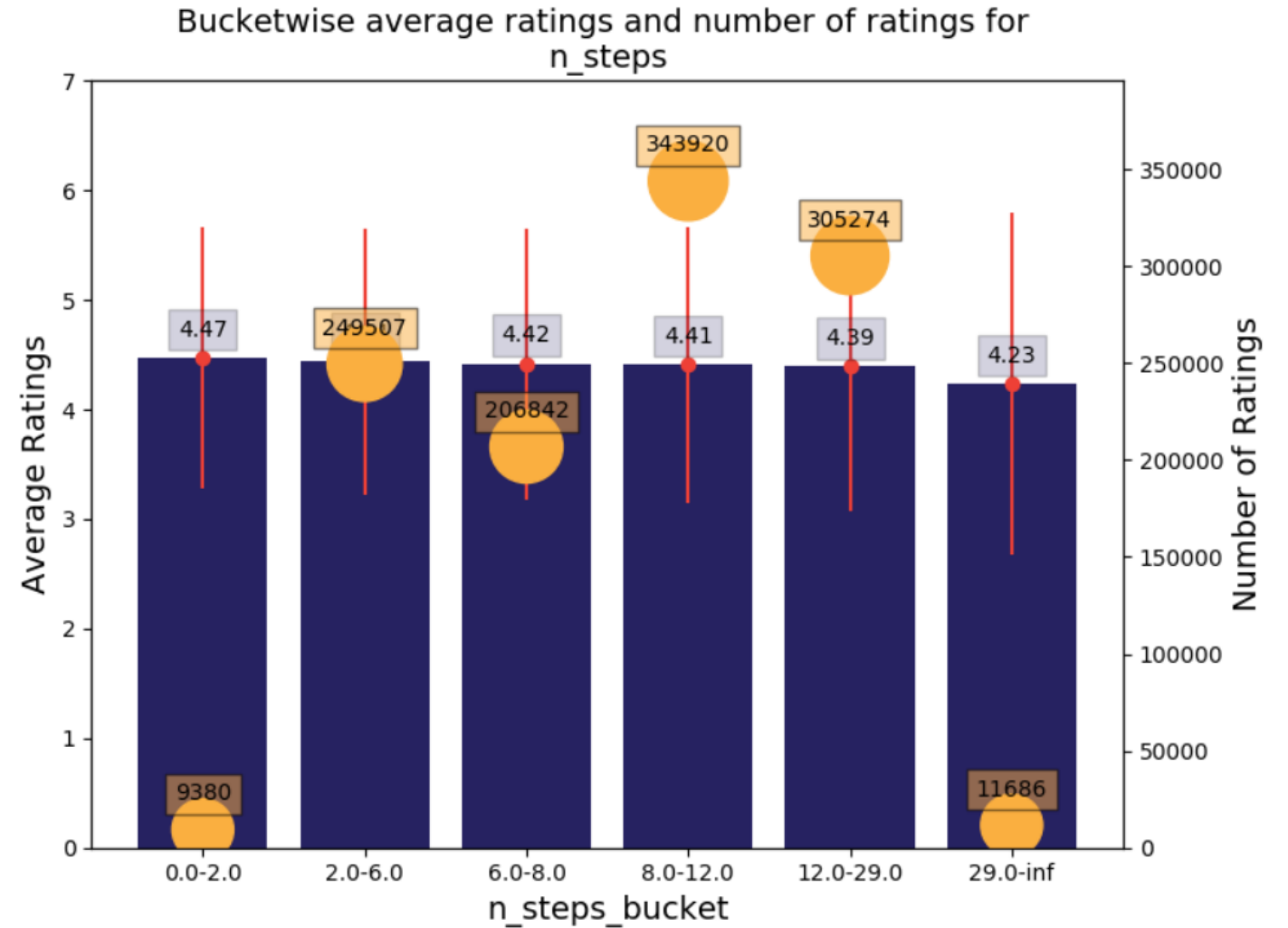
- ❖ Somewhat relevant
- ❖ Low prep time is preferred



EDA

Number of steps

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

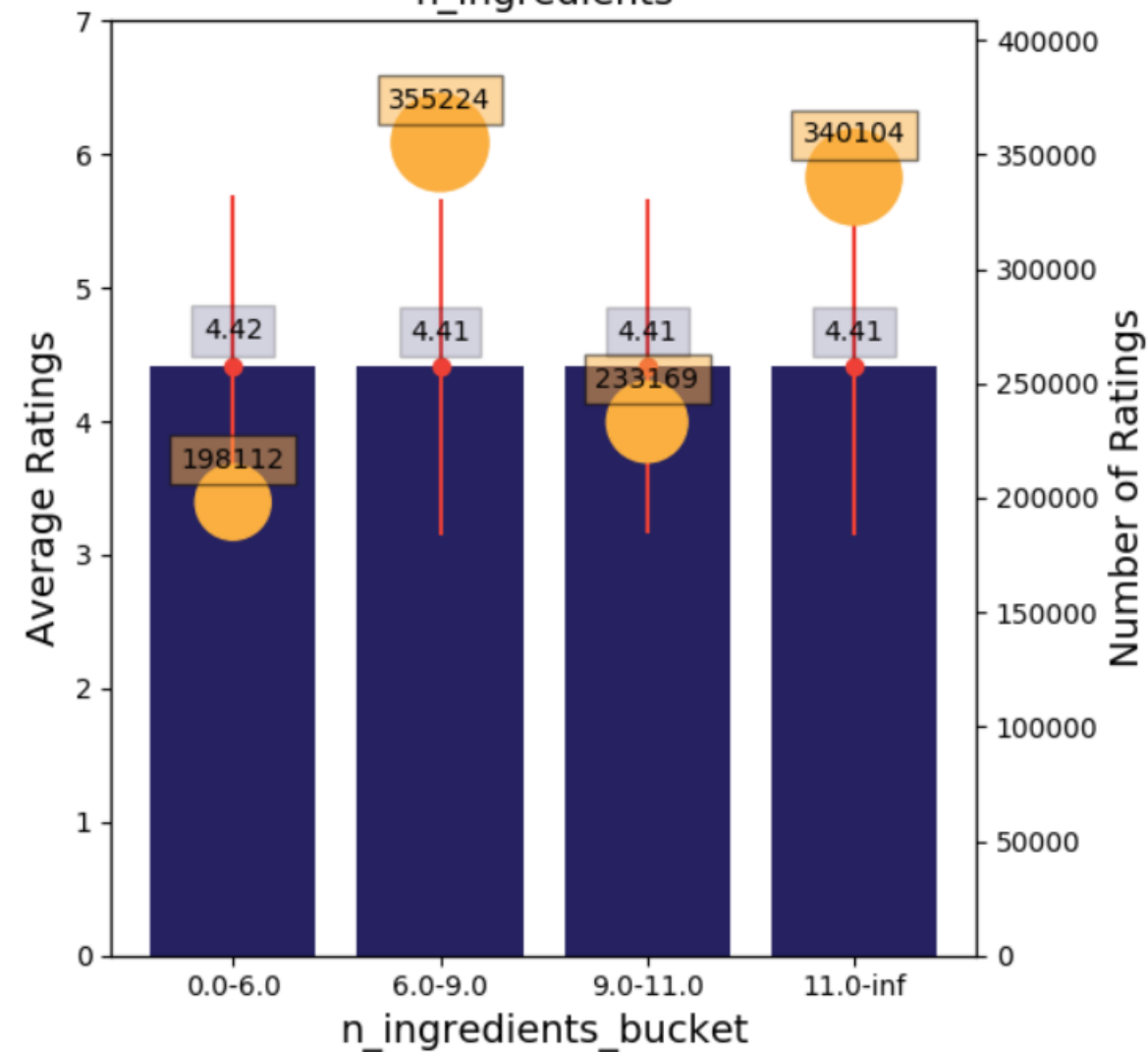


EDA

Number of ingredients

❖ Not relevant

Bucketwise average ratings and number of ratings for
n_ingredients



EDA

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

EDA

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

EDA

Top numbers most rated tags

individual_tag
preparation
time-to-make
course
dietary
main-ingredient
easy
occasion
equipment
cuisine
low-in-something
main-dish
60-minutes-or-less
number-of-servings
meat
taste-mood
north-american
30-minutes-or-less
vegetables
oven
4-hours-or-less

EDA

Bottom number least rated tags

```
+-----+
|      individual_tag|
+-----+
|      cranberry-sauce|
|           pot-roast|
|  main-dish-seafood|
|  ham-and-bean-soup|
|lamb-sheep-main-dish|
+-----+
```

EDA

Top number rated tags

```
+-----+
|      individual_tag|
+-----+
|  side-dishes-beans|
|                cabbage|
|heirloom-historic...|
|middle-eastern-ma...|
|  breakfast-potatoes|
+-----+
```


CONCLUSION and RECOMMENDATIONS

- ❖ In this analysis of recipe data shows that the review time since submission and the number of steps, preparation time and number of ingredients are important factors in determining the rating of a recipe.
- ❖ Recipes that are reviewed by users after a long time from the submission date, have less number of steps, less preparation time and less number of ingredients tend to have high ratings of 5.
- ❖ In contrast, the number of ingredients in a recipe is not found to be relevant to the rating. Similarly, the nutrition columns such as calories, fat, sugar, sodium, protein, and fat. and per serving are not found to be relevant in determining the rating of a recipe.
- ❖ The findings of this analysis can be used to inform decisions about recipe development and presentation to users in order to meet their preferences.