

San José State University Department of Applied Data Science

DATA 225

Database Systems for Analytics

Instructor: Simon Shim

Group Project

Cloud Analytics and Data
Warehouse Implementation

Group 5

Group Members:

Sowmya Neela

Naga Shreya Chillumukuru

Tanmay Singh

Terrence Wang

Neha Mrutyunjaya
Bhadragoudar

TABLE OF CONTENTS

- Abstract
- Dataset Description
- Whole cloud architecture
- Cloud SQL/NOSQL DB schema
- ETL/ELT processes and justifications
- Airflow pipeline
- DW schema and implementations
- Business Intelligence/Visualization/Data Analytics
- Analysis and Recommendations
- Conclusion

Abstract

This project is based on a digital brand and online social networking service featuring recipes from home cooks and celebrity chefs, Food.com. The website provides various recipes and reviews from netizens. We use historical data to analyze what kind of food people like, based on category, preparation time, instructions and nutrients such as protein, fat and sugar. Build models to recommend recipes and make predictions about recipes that people will like in the future based on trends. In addition, we will also obtain trending data from various websites in real time to verify and improve our models.

1. Dataset Description

1. Culinary Trends:

Analysis of the 'foodporn' dataset could reveal current culinary trends, popular ingredients, and emerging food preferences within the online community.

2. Cultural and Regional Insights:

- The dataset might showcase a diverse range of cuisines, reflecting global and regional food preferences. This could provide insights into cultural influences on food choices.

3. User Engagement and Interaction:

- Understanding which food posts receive the most engagement (likes, comments) can provide insights into user preferences and the types of content that resonate with the community.

4. Content Creation Patterns:

- Analysis of posting patterns and frequencies could shed light on when users are most active and what prompts them to share food-related content. This information can be valuable for content creators and marketers.

5. Influencer Impact:

- Identification of influential users or contributors within the 'foodporn' community can be useful for brands and marketers looking to collaborate with key influencers in the food industry.

6. Seasonal and Event-Driven Patterns:

- Analysis may reveal seasonal variations in food preferences and the impact of major events (holidays, food festivals) on the types of food content shared.

7. Recipe and Dish Popularity:

- The dataset could highlight popular recipes, dishes, and cooking techniques, providing insights into the foods that capture the community's attention.

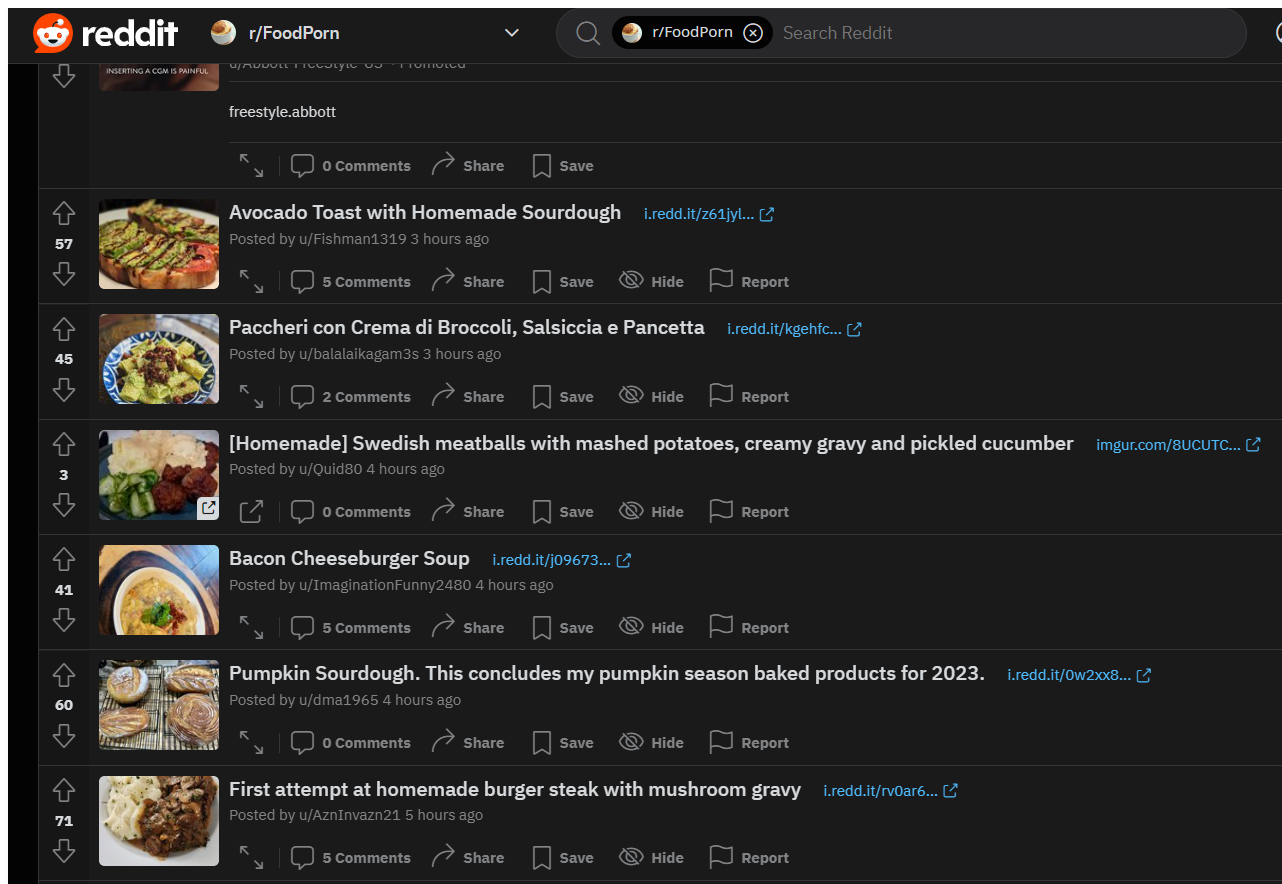
8. Potential for Recommendation Systems:

- Patterns in user preferences could be leveraged for the development of recommendation systems, suggesting similar content to users based on their engagement history.

Archival dataset:

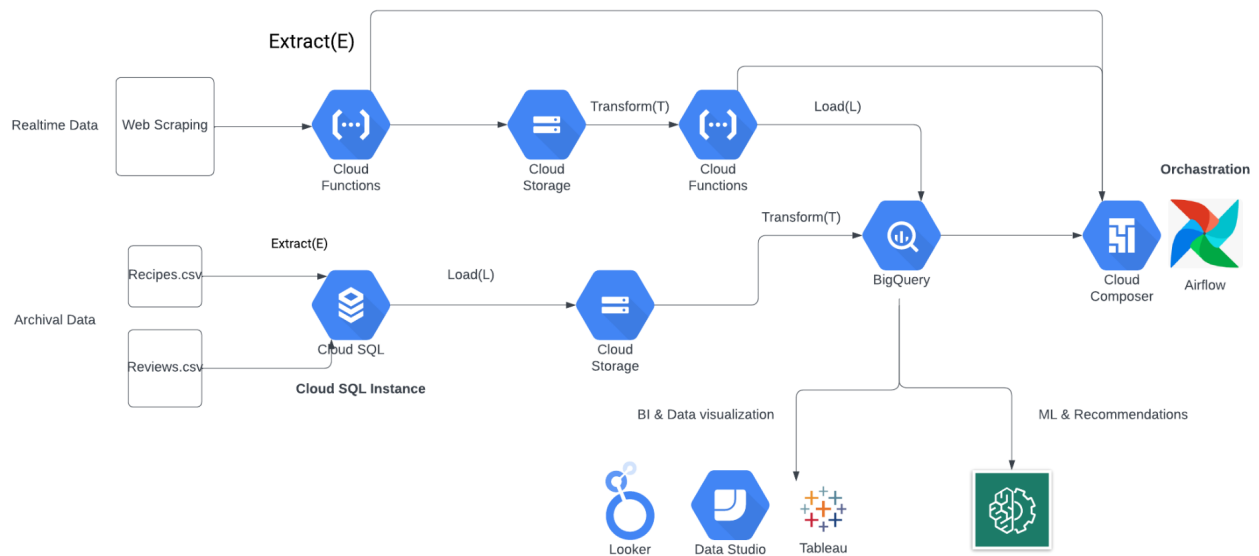
The dataset contains 522,517 recipes from 312 different categories. The reviews dataset contains 1,401,982 reviews from 271,907 different users. This dataset provides information about the author, rating, review text, and more.

Real-time dataset:



This dataset contains recipe names and the date on which they were posted on Reddit. This dataset is updated daily at midnight UTC. On average 100 titles per day are expected to be loaded in this dataset by our Praw (Python Reddit API Wrapper) API.

2. Whole cloud architecture



All our archival raw data is stored in GCP's Cloud SQL instance, loaded into a Cloud Storage bucket. Using GCP's data warehouse BigQuery, we extract and process the data from the bucket. The entire ELT (Extract, Load, Transform) process is scheduled and automated through Airflow's DAG (Directed Acyclic Graph) script. The processed data is further utilized for data visualization based on requirements, as well as for training machine learning models and implementing a recommendation system.

The process involves using Praw, a Python library, to collect today's post titles from food-related subreddits and then employing the OpenAI API to distill these titles into recipe names, which are saved as CSV files in cloud storage. A cloud function then retrieves the latest CSV file from each subreddit, adding a 'created_dt' field to mark the recipe's addition date and a 'column name' field for easy integration into BigQuery. Finally, this refined dataset is loaded into a BigQuery table named 'realtime_data_food', with the cloud function ensuring it augments the existing dataset.

3. Cloud SQL DB schema

We are using GCP's Cloud SQL instance to store the raw archive data. The instance is configured to have a private VPC network

The screenshot shows the Google Cloud console interface for configuring a Cloud SQL instance. The instance is named 'projectsql225' and is a MySQL 8.0 instance. The 'NETWORKING' tab is selected, showing the 'Instance IP assignment' section. In this section, the 'Private IP' checkbox is checked, indicating that the instance will use a private IP address. Below this, the 'Associated networking' section shows a dropdown menu for 'Network' set to 'default'. A green checkmark indicates that the private services access connection for the 'default' network has been successfully created. The 'Allocated IP range (optional)' section shows a dropdown menu for 'Allocated IP range' set to 'Use automatically assigned IP range'. Below this, the 'Public IP' checkbox is unchecked. The 'Authorized networks' section shows a dropdown menu for 'Terrence (192.168.1.110)' with a '(Not saved)' status. A red box highlights the 'Instance IP assignment' section, and another red box highlights the 'Authorized networks' section.

Google Cloud Data225

Connections

PRIM... All instances > projectsql225

projectsql225

MySQL 8.0

SUMMARY NETWORKING SECURITY CONNECTIVITY TESTS

Choose how you want your source to connect to this instance, then define which networks are authorized to connect. [Learn more](#)

You can use the Cloud SQL Proxy for extra security with either option. [Learn more](#)

Instance IP assignment

☒ Private IP

Assigns an internal, Google-hosted VPC IP address. Requires additional APIs and permissions. Can't be disabled once enabled. [Learn more](#)

Associated networking

Select a network to create a private connection

Network * default

Private services access connection for network **default** has been successfully created. You will now be able to use the same network across all your project's managed services. If you would like to change this connection, please visit the [Networking page](#).

Allocated IP range (optional)

Select an allocated IP range name to specify IP addresses your instance can connect with. Can't be changed after instance creation. [Learn more](#)

Allocated IP range Use automatically assigned IP range

[HIDE ALLOCATED IP RANGE OPTION](#)

☐ Public IP

Assigns an external, internet-accessible IP address. Requires using an authorized network or the Cloud SQL Proxy to connect to this instance. [Learn more](#)

Authorized networks

You can specify CIDR ranges to allow IP addresses in those ranges to access your instance. [Learn more](#)

Terrence (192.168.1.110) (Not saved) ▼

[ADD A NETWORK](#)

Archive raw dataset in Cloud SQL:

We have two tables for the archive data, Recipes and Review tables, 'RecipeId' is the primary key in Recipes table, and it is also the foreign key in Reviews table to join two tables.

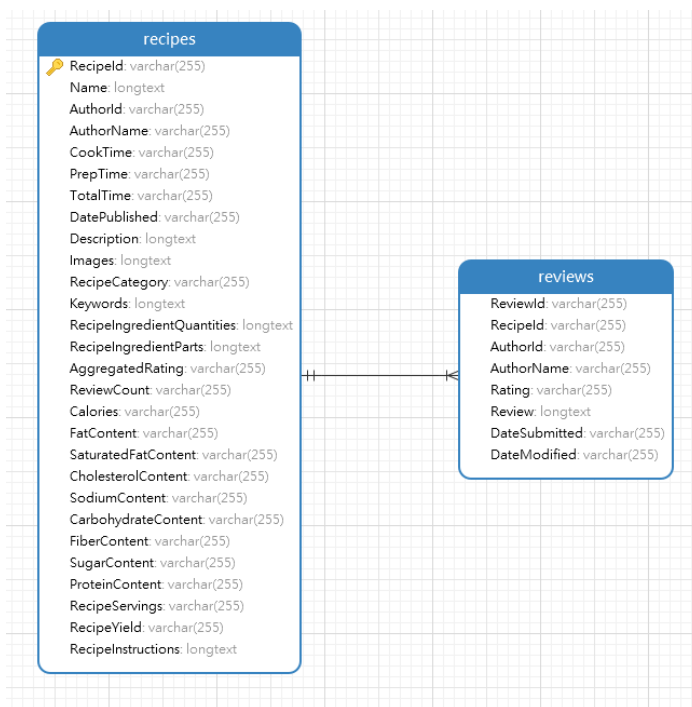
Recipes:

- **RecipeId:** The id of the recipe, primary key
- **Name:** The name of the recipe
- **AuthorId:** The id of the author of the recipe
- **AuthorName:** The name of the author of the recipe
- **CookTime:** The cooking time required for the recipe
- **PrepTime:** The preparation time required for the recipe
- **TotalTime:** The total time required for the recipe
- **DatePublished:** The date when the recipe was published
- **Description:** A description of the recipe
- **Images:** Images related to the recipe
- **RecipeCategory:** The category to which the recipe belongs
- **Keywords:** Keywords associated with the recipe
- **RecipeIngredientQuantities:** Quantities of ingredients used in the recipe
- **RecipeIngredientParts:** Parts of ingredients used in the recipe
- **AggregatedRating:** The aggregated rating of the recipe
- **ReviewCount:** The count of reviews for the recipe
- **Calories:** The calorie content of the recipe
- **FatContent:** The fat content of the recipe
- **SaturatedFatContent:** The saturated fat content of the recipe
- **CholesterolContent:** The cholesterol content of the recipe
- **SodiumContent:** The sodium content of the recipe
- **CarbohydrateContent:** The carbohydrate content of the recipe
- **FiberContent:** The fiber content of the recipe
- **SugarContent:** The sugar content of the recipe
- **ProteinContent:** The protein content of the recipe

- **RecipeServings:** The number of servings the recipe provides
- **RecipeYield:** The yield of the recipe
- **RecipeInstructions:** Instructions for preparing the recipe

Reviews:

- **ReviewId:** The id of the review, primary key
- **RecipeId:** The id of the recipe to which the review belongs, foreign key referencing RecipeId in Recipes table
- **AuthorId:** The id of the author of the review
- **AuthorName:** The name of the author of the review
- **Rating:** The rating given in the review
- **Review:** The content of the review
- **DateSubmitted:** The date when the review was submitted
- **DateModified:** The date when the review was last modified



Realtime dataset in Cloud SQL:

We have two tables for the realtime data in one table.

realtime_data_food:

- **Dish Name:** Name of the dish
- **Created_dt:** Source file created date.

4. ETL/ELT processes and justifications

Archive data:

We are performing the Extract(E), Load(L), Transform(T) process for the archive dataset. We initially upload the raw data to GCP **Cloud Storage** and import it into **Cloud SQL** storage (E). Subsequently, we load the data into the **BigQuery** data warehouse (L) for transformation processes (T), which include cleaning unnecessary redundant fields such as 'Description,' 'Images,' 'RecipeYield,' and 'RecipeInstructions.'

Moreover, in the original dataset, CookTime, PrepTime and TotalTime are stored as non-standard value types.

	RecipeId	Name	AuthorId	AuthorName	CookTime	PrepTime	TotalTime
0	38	Low-Fat Berry Blue Frozen Dessert	1533	Dancer	PT24H	PT45M	PT24H45M
1	39	Biryani	1567	elly9812	PT25M	PT4H	PT4H25M

We unify them into minutes and store them as integers.

CookTimeInMinutes	PrepTimeInMinutes	TotalTimeInMinutes
1440	45	1485
25	240	265
5	30	35

Real-Time :

We are performing the Extract(E), Transform(T), and Load(L) processes for this dataset.

Extract(E)

There are subreddits where people post about the food they are eating or making. We need to extract the titles of all the posts that were posted today.

on FoodPorn people are posting about recipe

We are connecting to Reddit using Praw to fetch titles from each subreddit after fetching the title we filter posts based on the latest date. After getting the title OpenAI API cleans the title and takes out only the recipe name and these names are saved as CSV files in cloud storage.

Transform (T)

So each subreddit's latest file is picked using the cloud function and a created_dt column is added to know on which date a certain recipe was loaded and the column name is added so that this data can be loaded into big query

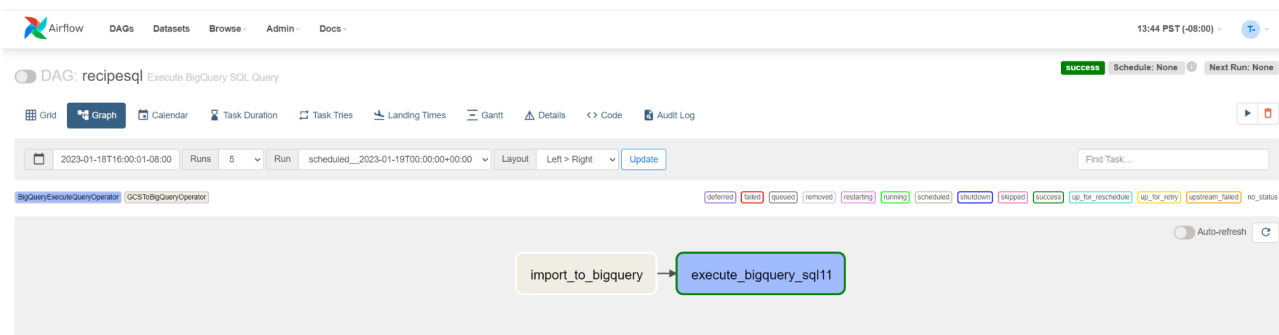
Load (L)

After the cleaned data frame is created we are going to load it into a big query table realtime_data_food using cloud function. This is going to be inserted and appended with existing data.

5. Airflow pipeline

The ETL pipeline is seamlessly orchestrated through Airflow's DAG script architecture, with Cloud Composer on GCP serving as the managed Apache Airflow service. DAG files efficiently design and schedule workflows, enhancing the project's workflow management. Cloud Composer's integration ensures streamlined execution of tasks, optimizing the overall efficiency of the data processing pipeline.

Archive Data:



For archive data, we have a DAG script to import dataset from Cloud Storage to Big Query data warehouse, then execute a Big Query SQL to clean the dataset. Airflow schedules the work flow.

The screenshot shows the Airflow web interface displaying a list of DAGs. The top navigation bar includes 'Airflow', 'DAGs', 'Datasets', 'Browse', 'Admin', and 'Docs'. The 'DAGs' tab is selected. The table below lists the DAGs:

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
airflow_monitoring	airflow	40	15 * * * *	2023-12-04, 13:30:00	2023-12-04, 13:40:00	1	▶ 🛑	...
recipesql	tanys.wang@cloud.edu	1	None	2023-01-18, 16:00:00		1	▶ 🛑	...

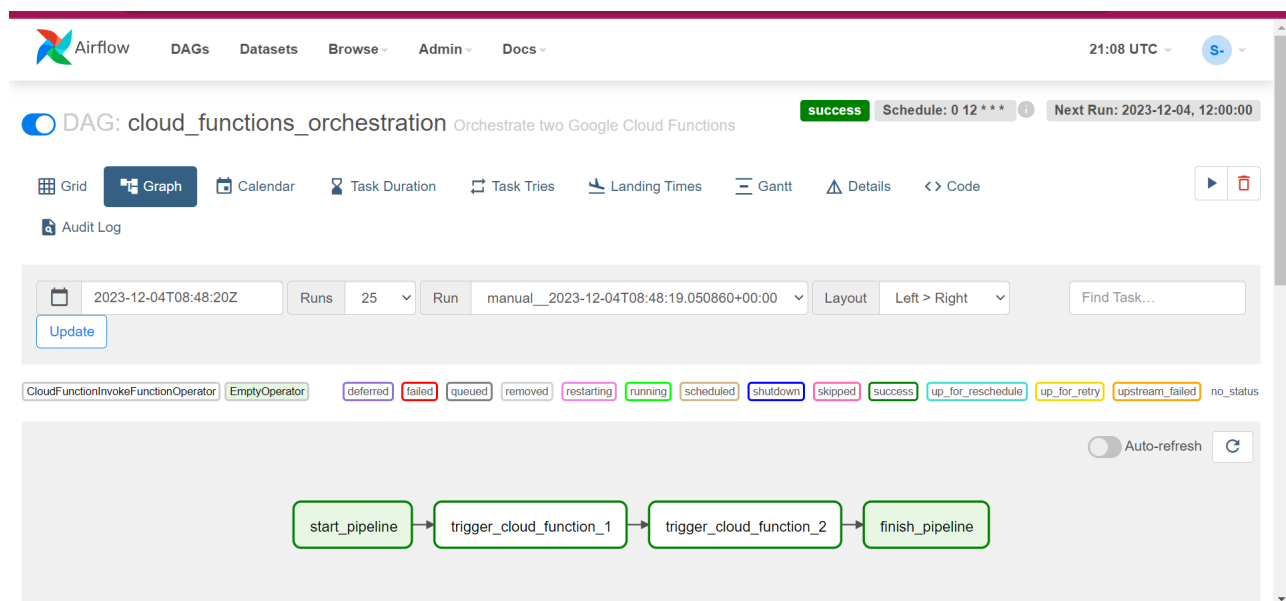
Realtime Data:

- The `'trigger_cloud_function_1'`, activates the `'web_scraping'` function, and is responsible for the extraction part of the ETL. This function uses Praw (a Reddit API Python wrapper) to scrape recent post titles from various food-related subreddits. This step focuses on collecting current date data, especially recipe names, which are processed and stored as CSV files in cloud storage.

- The `'trigger_cloud_function_2'`, triggering the `'load_file_from_bucket_to_bigquery'` function, handles the transformation phase. Here, the newest CSV file from each subreddit is selected. The data is then enriched by adding fields like `'created_dt'` for easier integration into BigQuery.

- Transforming the data ensures it's ready for analysis and organized storage in a database like BigQuery.

In essence, the Airflow DAG is designed to streamline this ETL process, where Reddit data is extracted, refined, and then loaded into BigQuery for efficient storage and analysis. This setup is key for handling data workflows quickly and automatically, essential for tasks involving real-time data processing.





realtime

All 2 Active 2 Paused 0

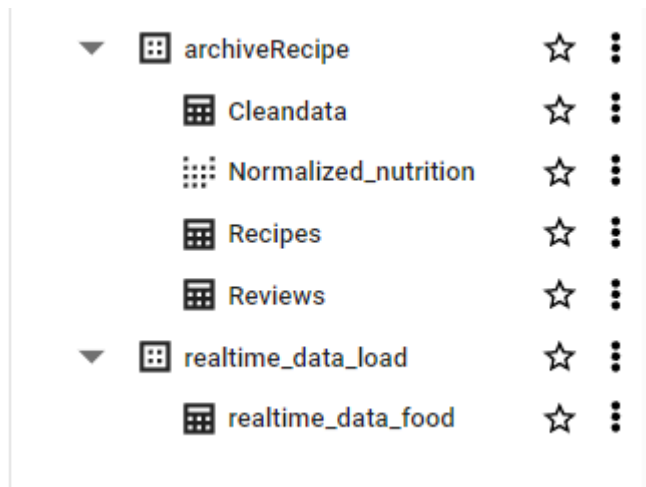
Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
airflow_monitoring	airflow	139	* / 10 * * * *	2023-12-04, 21:20:00	2023-12-04, 21:30:00	1
cloud_functions_orchestration	sowmya.neela@sjsu.edu	4	0 12 * * *	2023-12-04, 08:48:19	2023-12-04, 12:00:00	4

6. DW schema and implementations



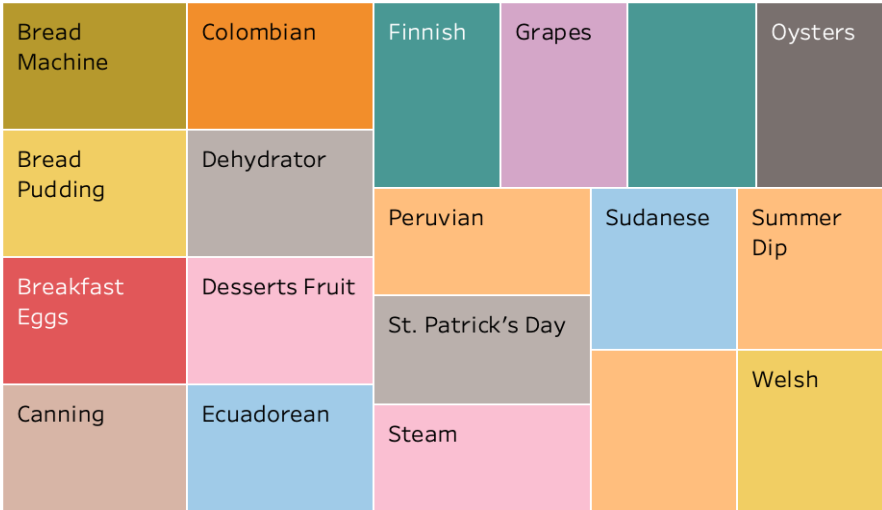
▼	archiveRecipe	☆	⋮
	Cleandata	☆	⋮
	Normalized_nutrition	☆	⋮
	Recipes	☆	⋮
	Reviews	☆	⋮
▼	realtime_data_load	☆	⋮
	realtime_data_food	☆	⋮

In our BigQuery environment, we manage two distinct datasets: 'archiveRecipe' and 'realtime_data_load.' Within each dataset, we maintain two primary tables, namely 'Recipes' and 'Reviews.' Notably, the 'Reviews' table is designed with a foreign key relationship that links to the 'Recipes' table, establishing a connection between the two.

In the course of our Extract, Load, Transform (ELT) process, we introduce a key intermediary table named 'Cleandata.' This table represents the outcome of our data cleansing procedures applied to the 'Recipes' table, resulting in a refined and sanitized dataset.

Additionally, we have a specialized table named 'Normalized_nutrition,' wherein all nutritional columns associated with recipes have undergone a normalization process. This table provides a standardized view of nutritional data for enhanced analytical insights.

Sheet 6



Recipe Category

- Bread Machine
- Bread Pudding
- Breakfast Eggs
- Canning
- Colombian
- Dehydrator
- Desserts Fruit
- Ecuadorean
- Finnish
- Grapes
- Memorial Day
- Oysters
- Peruvian
- St. Patrick's Day
- Steam
- Sudanese
- Summer Dip
- Venezuelan
- Welsh

Name

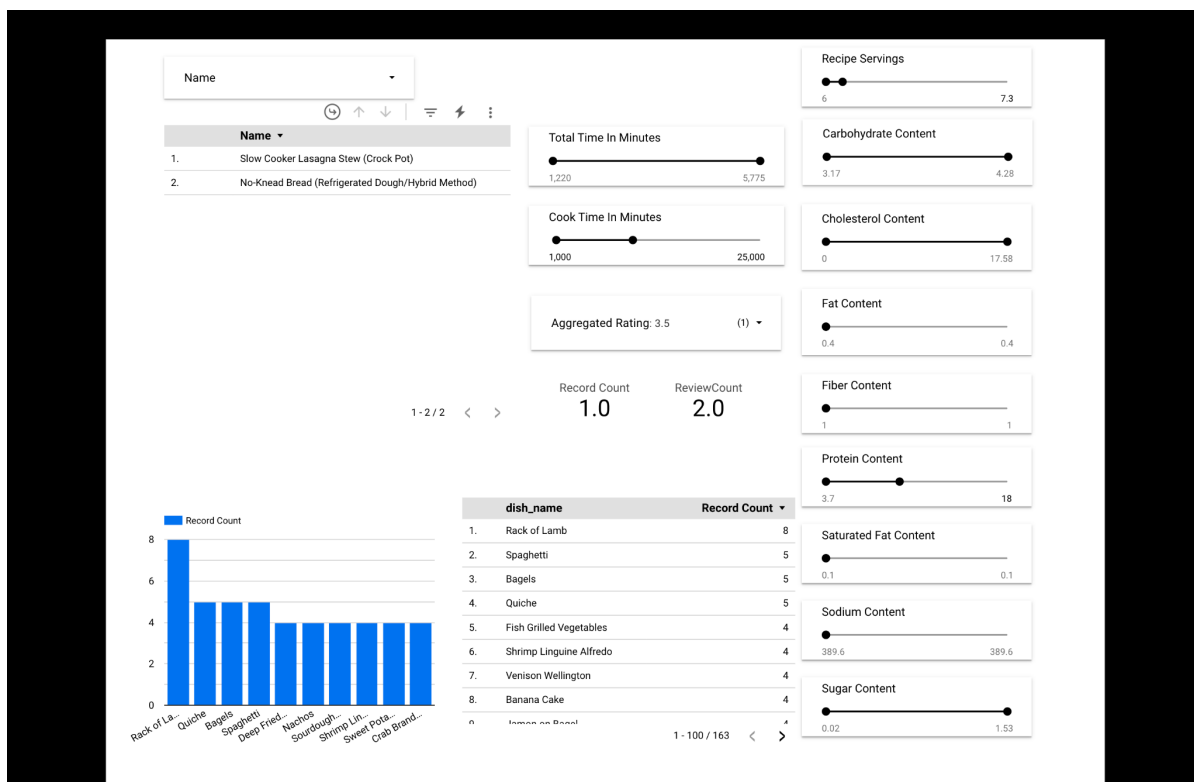
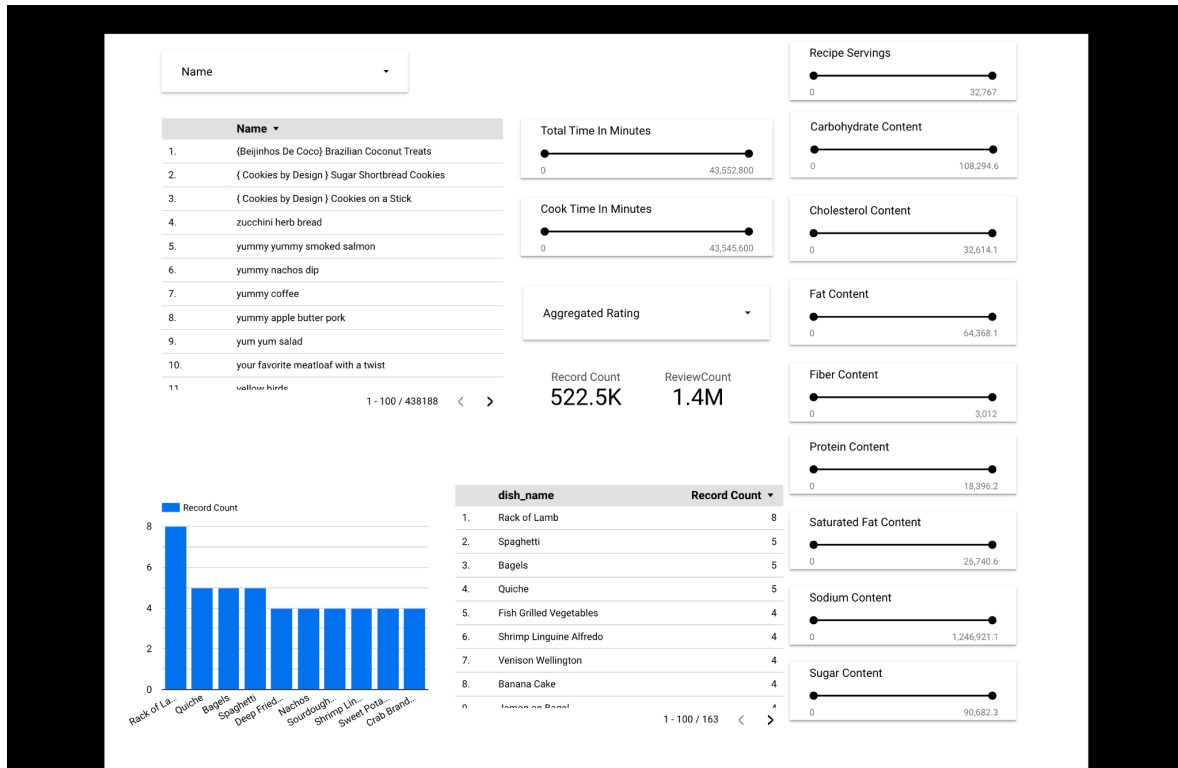
- "Better Than Sex" S..
- "Backyard-Style" Ba..
- "A Bit Different" Bre..
- ...from Stephanie's K..
- V's Kicked up Bake..
- Light-- Berry Loaf
- Baked Potato-- Bak..
- Hawaiian Sunr..
- Tasty's -- the Powe..
- Tasty Dish's -- Not..

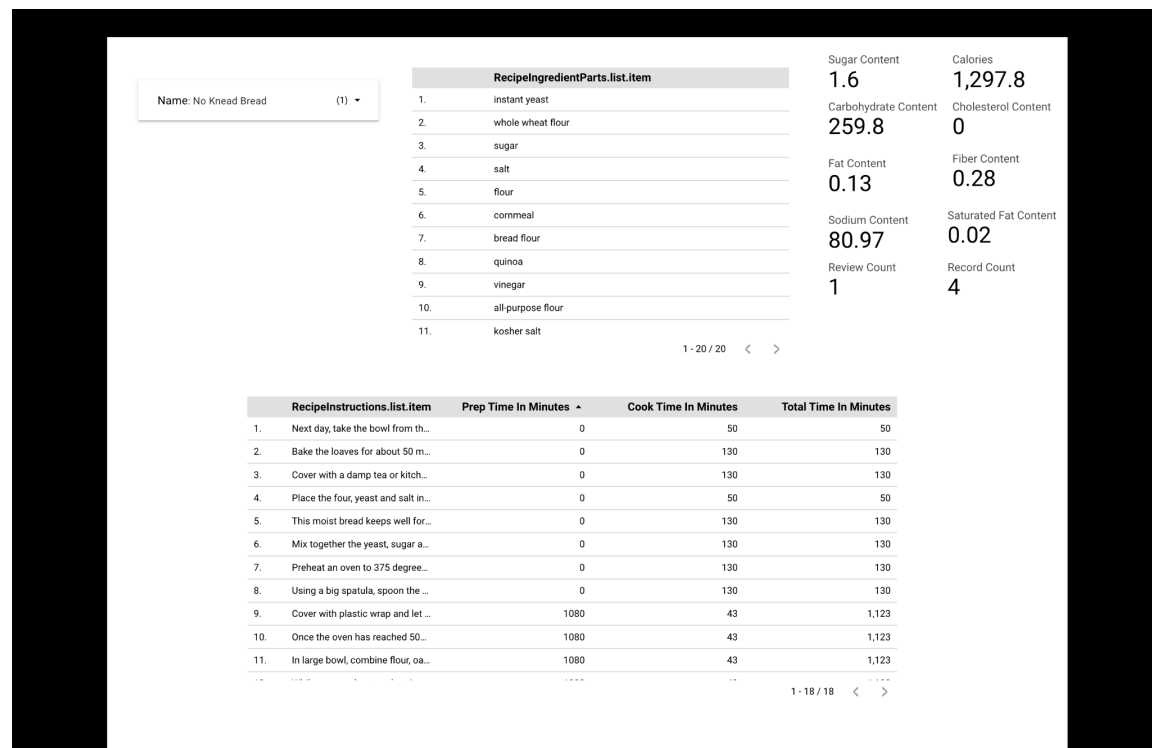
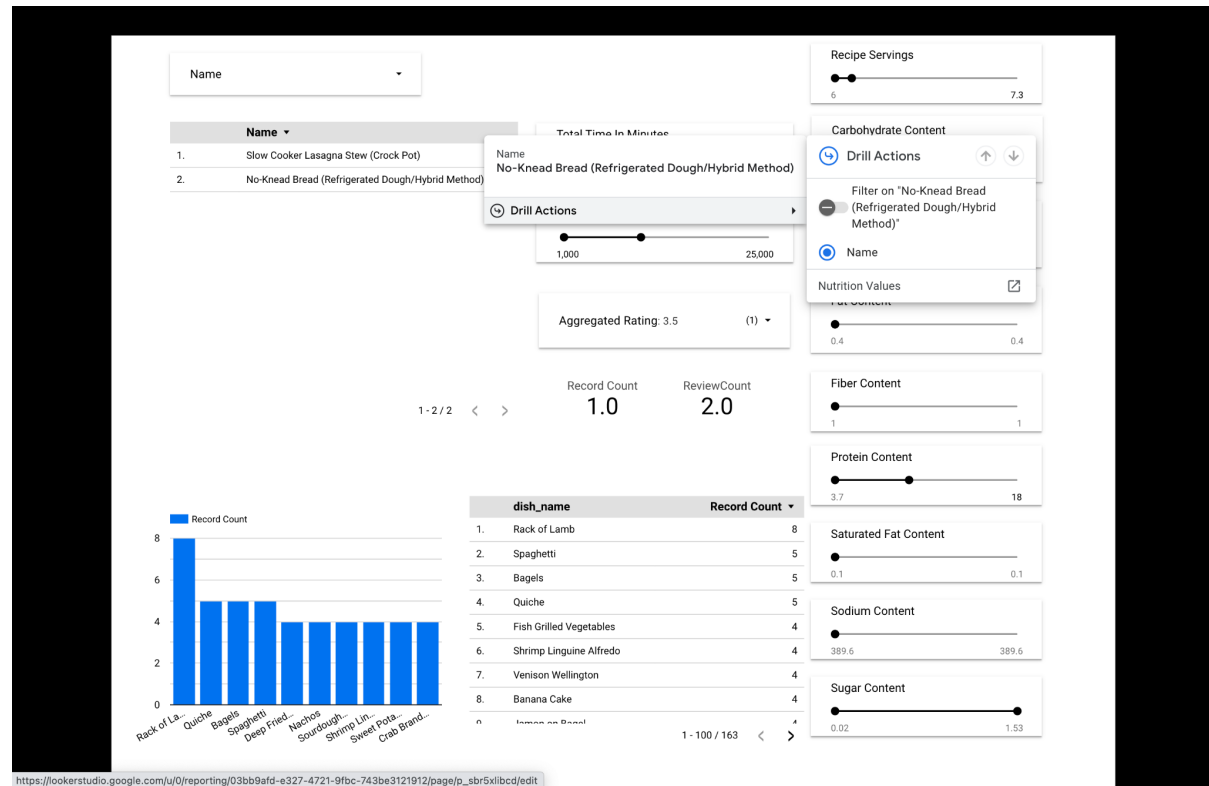


Visualizations made on Looker Studio

This dashboard shows the “Name” dropdown, the sliders for various nutritional values along with Recipe

<https://lookerstudio.google.com/u/0/reporting/03bb9afd-e327-4721-9fbc-743be3121912/page/xi2jD/edit>





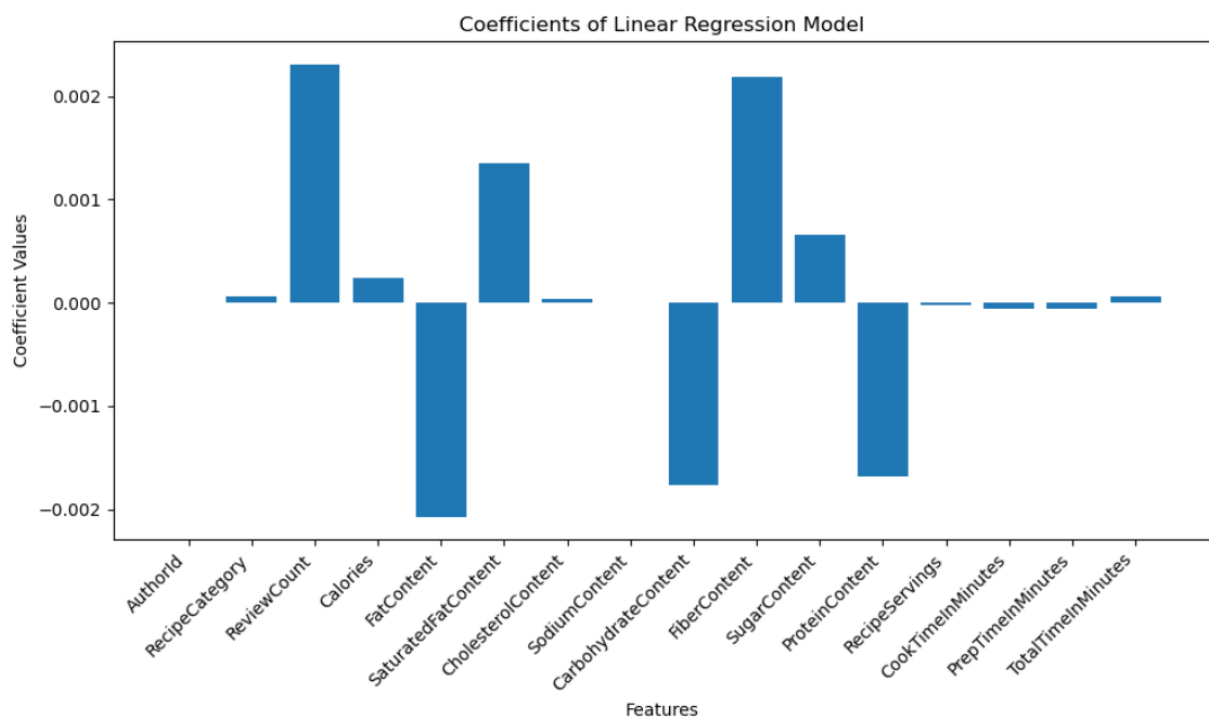
8. Analysis and Recommendations

Regression analysis:

We conducted a Linear Regression analysis with the Recipe Rating as the dependent variable and other features as independent variables. The coefficients obtained from this analysis provided valuable information about the strength and direction of the impact each feature has on the recipe rating. Key coefficients revealed the relative importance of different attributes.

```
# Show coefficients for each feature
for feature, coef in zip(X_train.columns, coefficients):
    print(f"{feature}: {coef}")
```

```
AuthorId: -5.949815767220795e-11
RecipeCategory: 5.713118945095987e-05
ReviewCount: 0.0023093414545950494
Calories: 0.00024143074067568116
FatContent: -0.0020718536498796846
SaturatedFatContent: 0.001351501184700554
CholesterolContent: 3.730046604255955e-05
SodiumContent: -5.271321261939683e-07
CarbohydrateContent: -0.0017649372593535983
FiberContent: 0.00218984545306187
SugarContent: 0.0006626881284555628
ProteinContent: -0.0016848685596623642
RecipeServings: -2.5586062879287404e-05
CookTimeInMinutes: -5.607920153747461e-05
PrepTimeInMinutes: -5.4689969509788766e-05
TotalTimeInMinutes: 5.626540255776237e-05
```



Features with positive coefficients such as RecipeCategory indicate a positive relationship with recipe ratings. An increase in these features tends to lead to higher ratings. Features with negative coefficients like SodiumContent suggest a negative association with recipe ratings. A decrease in these features may result in higher ratings. At the same time, we also found that recipes are greatly affected by publishers and have certain content orientation and fan orientation.

Classification analysis

Trained a Correlation model using nutrition values.

Displays the top 10 similar recipes with nutrition values by taking the Recipe ID as input, and having the nutrition values of that recipe ID as reference.

for our recipe Carrot Cake

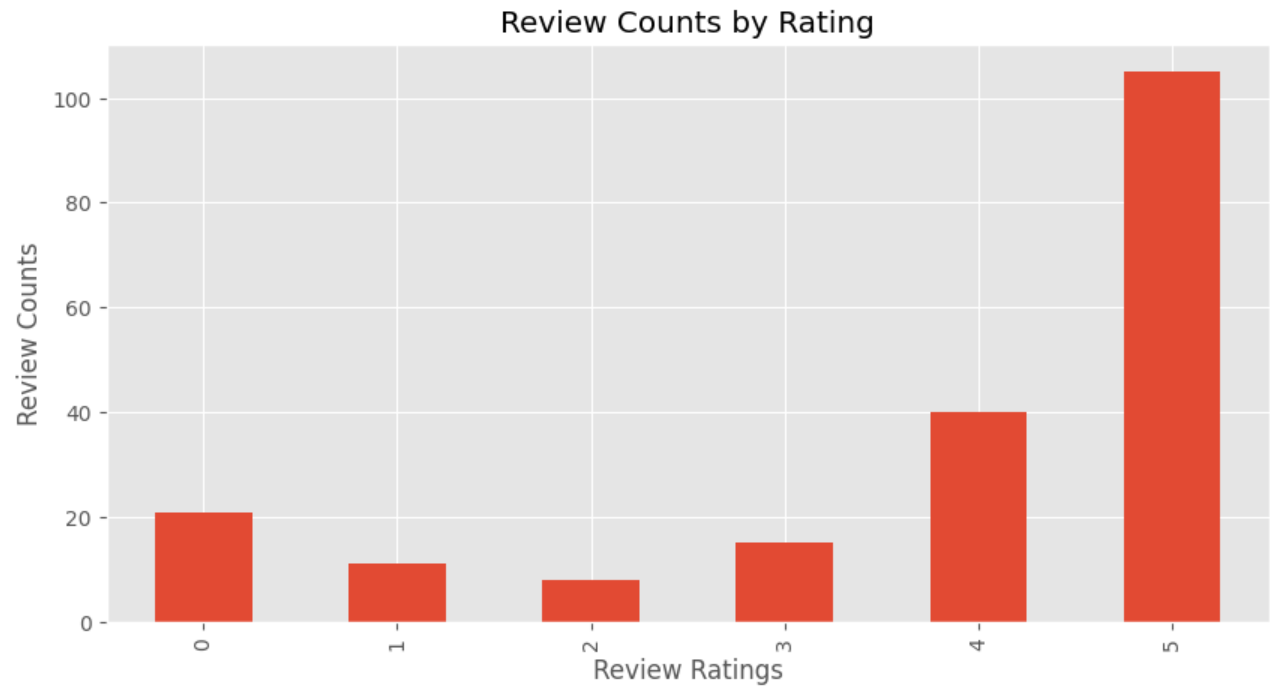
[56] ✓ 0.0s

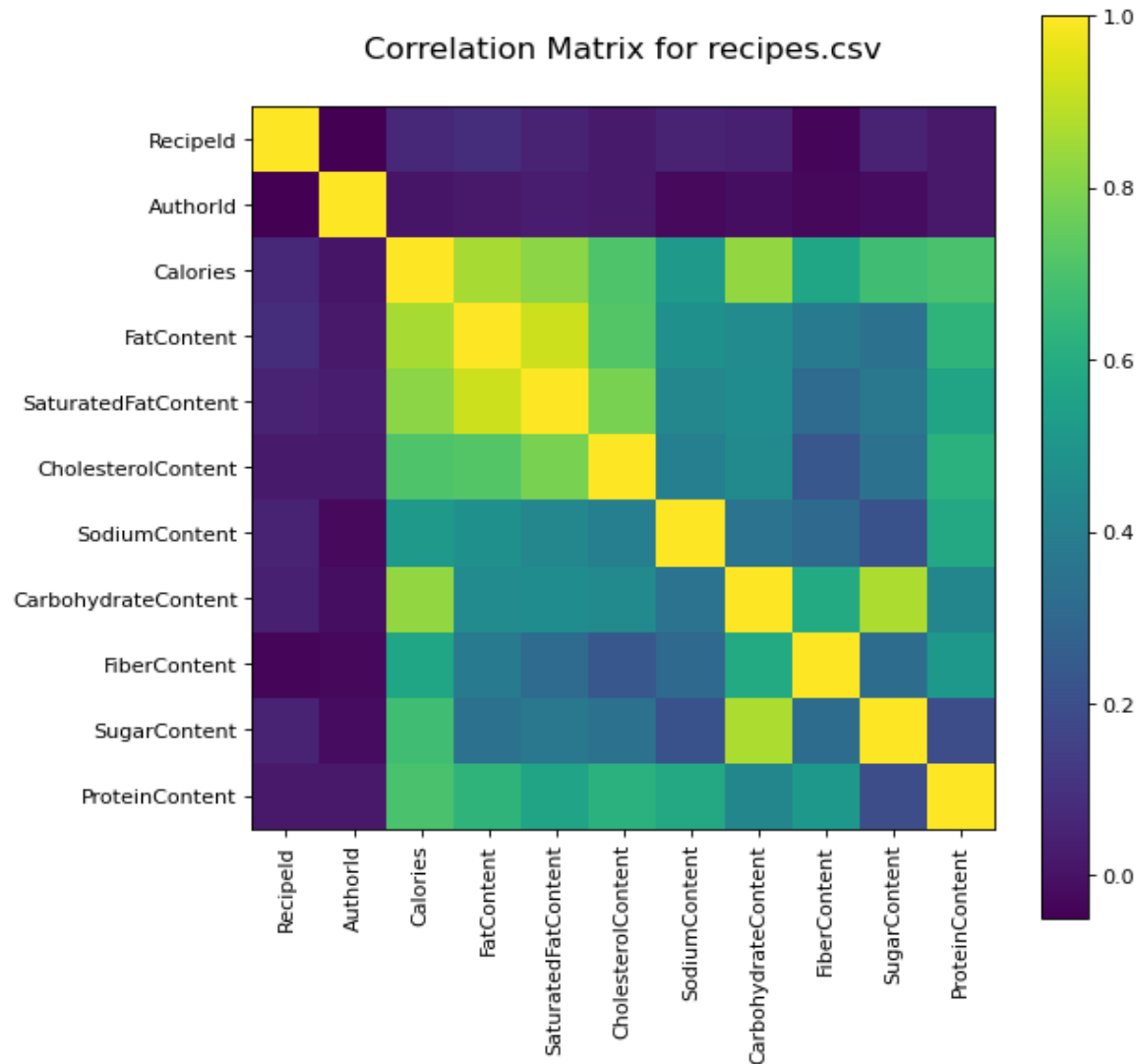
Python

RecipeId	Name	Calories	FatContent	SaturatedFatContent	CholesterolContent	SodiumContent	CarbohydrateContent	FiberContent	SugarContent	ProteinContent	RecipeServings
16	54.0 Carrot Cake	0.0001	0.001061	0.000425	0.000178	0.000063	0.000052	0.000076	0.000044	0.00021	12.0

These are the similar recipes to it

	RecipeId	Name	Calories	FatContent	SaturatedFatContent	CholesterolContent	SodiumContent	CarbohydrateContent	FiberContent	SugarContent	ProteinContent	RecipeServings
472229	489638.0	Storm Warning Glazed Cinnamon Coffee Cake	0.000103	0.001073	0.000425	0.000189	0.000039	0.000054	0.000057	0.000049	0.000227	12.0
443278	459624.0	Strawberry Cake	0.000112	0.001065	0.000434	0.000161	0.000042	0.000063	0.000062	0.000058	0.000202	12.0
72397	77150.0	Traditional Black Russian Bundt Cake	0.000098	0.001049	0.000382	0.000158	0.000058	0.000044	0.000062	0.000037	0.000206	12.0
182508	190783.0	Ho Ho Ho Rum Cake	0.000102	0.001006	0.000416	0.000182	0.000059	0.000051	0.000057	0.000045	0.000202	12.0
218293	227581.0	Cranberry Apple Cake	0.000105	0.001069	0.000442	0.000199	0.000039	0.000055	0.000124	0.000041	0.000236	12.0
61564	65963.0	Lemon Key Lime Cake	0.000096	0.001053	0.000425	0.000195	0.000050	0.000048	0.000024	0.000036	0.000172	12.0
401614	416338.0	Irish Cream Almond Cake	0.000103	0.001088	0.000468	0.000204	0.000059	0.000050	0.000062	0.000044	0.000248	12.0
18730	22110.0	Sock-It-To-Me Cake	0.000088	0.001030	0.000477	0.000204	0.000037	0.000040	0.000076	0.000032	0.000231	12.0
403283	418057.0	Carrot Cake With Baby Food Carrots	0.000118	0.001038	0.000477	0.000212	0.000048	0.000069	0.000062	0.000066	0.000236	12.0
401653	416380.0	Pantry Cake	0.000104	0.000994	0.000399	0.000201	0.000039	0.000058	0.000067	0.000053	0.000227	12.0





9. Conclusion

In conclusion, our dashboard represents a comprehensive approach to recipe exploration, offering users the flexibility to filter recipes based on diverse criteria such as nutrition values, review count, and ratings. The integration of a KNN model enhances the user experience by presenting the top 10 closest recipes, strategically leveraging the similarity among our diverse collection.

Moreover, our advanced model considers a multitude of factors, including calories, nutrition values, prep time, and total cooking time, to unveil the intricate correlations influencing recipe ratings. This sophisticated analysis provides a holistic perspective, enabling users to discover not only the most similar recipes but also those that align closely with their specific preferences and dietary considerations. Through these innovative features, our dashboard aims to elevate the culinary exploration experience, catering to the varied tastes and preferences of our users. We've incorporated a feature that taps into real-time data to identify trending recipes. By analyzing which

recipes are repeatedly posted and enjoyed by many people, this keeps us up-to-date with popular and buzzworthy recipes.

10. **References**

- <https://ieeexplore.ieee.org/document/8765311>
- <https://arxiv.org/pdf/1905.06269>
- <https://github.com/Universe-89/Recipe-Radar> - GitHub link