Group X: Group project

Group members and their responsibilities

❑ Konstantin Benischke: Data transformation, presentation and discussion of results

❑ Milan Cukovic: Data ingestion and data handling

❑ Marie Nitschke: Research problem and motivation, data queries

❑ Fanni Tuominen: Research problem and motivation, data modeling and handling

Research question

As students, we are busy with assignments and the fridge is mostly empty, but we still want to eat healthy and tasty food so that we have energy for our tasks.

That is why we wanted to find out, if it is possible to find the best possible recipes with as little time and effort as possible.

Our research questions are:

1. Can one cook a good rated and healthy meal, while also trying to minimize the time needed?

1. Can one make a good rated and healthy meal, while following recipes with very few ingredients?

Link to the dataset:

https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions

❏ The provided dataset consists of two files containing the data to perfectly depict the relation between eg. rating of a recipe, amount of time and ingredients needed.

❏ To represent our database structure visually, we will use draw.io. Its user-friendly interface and adaptability make it an ideal platform for crafting an ER (Entity-Relationship) diagram that impeccably illustrates our database schema.

❏ Python will serve as our primary tool for transforming the raw data into a usable format and later generating graphical representations of our final results.

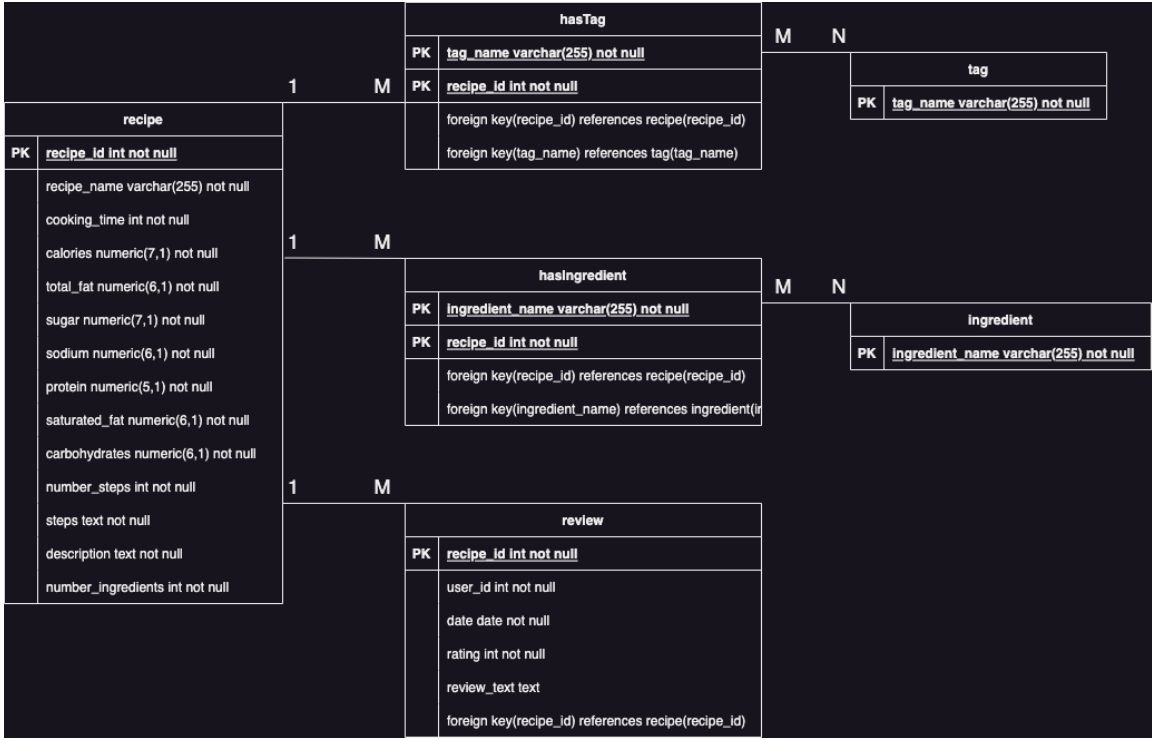
❏ In order to have effective data management, we've elected PostgreSQL over the conventional CSV files as our database management system. With this we ensure optimal handling and querying of large datasets while upholding the integrity of our data through a robust relational database structure.

Data modeling and data ingestion

Transforming and ingesting data with python and Jupyter Notebook

Used datasets:

- RAW_recipes.csv
- RAW_interactions.csv



recipe	
PK	recipe_id int not null
	recipe_name varchar(255) not null
	cooking_time int not null
	calories numeric(7,1) not null
	total_fat numeric(6,1) not null
	sugar numeric(7,1) not null
	sodium numeric(6,1) not null
	protein numeric(5,1) not null
	saturated_fat numeric(6,1) not null
	carbohydrates numeric(6,1) not null
	number_steps int not null
	steps text not null
	description text not null
	number_ingredients int not null

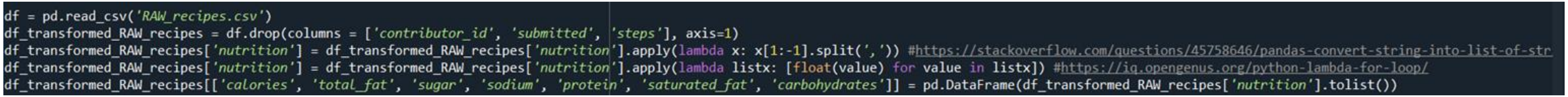
hasTag	
PK	tag_name varchar(255) not null
PK	recipe_id int not null
	foreign key(recipe_id) references recipe(recipe_id)
	foreign key(tag_name) references tag(tag_name)

tag	
PK	tag_name varchar(255) not null

hasIngredient	
PK	ingredient_name varchar(255) not null
PK	recipe_id int not null
	foreign key(recipe_id) references recipe(recipe_id)
	foreign key(ingredient_name) references ingredient(i

ingredient	
PK	ingredient_name varchar(255) not null

review	
PK	recipe_id int not null
	user_id int not null
	date date not null
	rating int not null
	review_text text
	foreign key(recipe_id) references recipe(recipe_id)



```
df = pd.read_csv('RAW_recipes.csv')
df_transformed_RAW_recipes = df.drop(columns = ['contributor_id', 'submitted', 'steps'], axis=1)
df_transformed_RAW_recipes['nutrition'] = df_transformed_RAW_recipes['nutrition'].apply(lambda x: x[1:-1].split(',')) #https://stackoverflow.com/questions/45758646/pandas-convert-string-into-list-of-str
df_transformed_RAW_recipes['nutrition'] = df_transformed_RAW_recipes['nutrition'].apply(lambda listx: [float(value) for value in listx]) #https://iq.opengenus.org/python-lambda-for-loop/
df_transformed_RAW_recipes[['calories', 'total_fat', 'sugar', 'sodium', 'protein', 'saturated_fat', 'carbohydrates']] = pd.DataFrame(df_transformed_RAW_recipes['nutrition'].tolist())
```

Database queries and data handling

Getting single recipes

```
SELECT DISTINCT r.*
FROM recipe r
JOIN review u ON r.recipe_id = u.recipe_id
JOIN hastag t ON r.recipe_id = t.recipe_id AND
t.tag_name = 'low-cholesterol'
JOIN hastag t2 ON r.recipe_id = t2.recipe_id AND
t2.tag_name = 'main-dish'  / “appetizers” /
“desserts”
WHERE u.rating >= 4.5
AND r.number_ingredients <= 5
AND r.sugar <= 100
AND r.total_fat <= 100
AND r.sodium <= 100
AND r.protein <= 100
AND r.saturated_fat <= 100
AND r.carbohydrates <= 100
AND r.cooking_time <= 15;
```

Getting recipe combinations

```
WITH FilteredRecipes AS (
  SELECT DISTINCT r.*
  FROM recipe r
  JOIN review u ON r.recipe_id = u.recipe_id
  JOIN hastag t ON r.recipe_id = t.recipe_id AND t.tag_name = 'low-cholesterol'
  WHERE u.rating >= 4.5
                              AND r.number_ingredients <= 5
                              AND r.sugar <= 100
    AND r.total_fat <= 100
    AND r.sodium <= 100
    AND r.protein <= 100
    AND r.saturated_fat <= 100
    AND r.carbohydrates <= 100
    AND r.cooking_time <= 15
)
SELECT DISTINCT main.recipe_id AS main_dish_id, main.recipe_name AS main_dish_name,
side.recipe_id AS sides_id, side.recipe_name AS sides_name
FROM FilteredRecipes main
CROSS JOIN FilteredRecipes side
JOIN hastag t_main ON main.recipe_id = t_main.recipe_id AND t_main.tag_name = 'main-dish'
JOIN hastag t_side ON side.recipe_id = t_side.recipe_id AND t_side.tag_name = “appetizers” /
'desserts'
WHERE main.number_ingredients + side.number_ingredients <= 10
  AND main.sugar + side.sugar <= 100
  AND main.total_fat + side.total_fat <= 100
  AND main.sodium + side.sodium <= 100
  AND main.protein + side.protein <= 100
  AND main.saturated_fat + side.saturated_fat <= 100
  AND main.carbohydrates + side.carbohydrates <= 100
  AND main.cooking_time + side.cooking_time <= 30;
```

Presentation and discussion of results
Recipes that meet our parameters

→ even after defining some hard parameters, we managed to find some recipes

Appetizers

- 181 out of 2553 (7,09%)
- Most common tags: (15-minutes-or-less, low-in-something,course, dietary, time-to-make, preparation)
- Most common ingredient: “salt”

Main-dishes

- 77 out of 7023 (1,1%)
- Most common tags: (15-minutes-or-less, low-in-something,course, dietary, time-to-make, preparation)
- Most common ingredient: “salt”

Desserts

- 314 out of 3581 (8,77%)
- Most common tags: (15-minutes-or-less, low-in-something,course, dietary, time-to-make, preparation)
- Most common ingredient: “sugar”

When asking for recipe combinations within our parameters of **main-dish** with either **appetizers** or **desserts,** we found *12.807 / 19.640* possible combinations

Reproducibility aspects

Link to the dataset:

https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions

1) Run “db_project_create_tables” to create the database
2) Run “data_transformator” to insert csv-data into the database
3) Run the queries with different parameters (main-dish, appetizers, desserts)
4) Run the “data_visualizer” file to get a visual representation

Used files:

- RAW_recipes.csv
- RAW_interactions.csv
- db_project_create_tables.sql
- data_transformator.ipynb
- data_visualizer.ipynb

Processed files:

- RAW_interactions_processed.csv
- RAW_recipes_processed.csv