CS/INFO 3300; INFO 5100

Project 3

1. Tasks Completed by Individuals

1.1. Emiri Aoki

- Integrated the HTML/CSS for recipe filtering with the hierarchical diagram
- Fixed issues with the diagram changing based on meal category and hovering over an ingredient to show ingredient relationships for that specific ingredient
- Contributed to final report -- wrote the Story part and building the visualization part
- Wrote descriptions and instructions for end user to use the graph correctly
- Hid left panel so it would show up only when a user hovers over ingredient
- Added additional visual elements to recipes including identifying which ingredient was hovered over last, and added colors to recipes to highlight what they are in the pie graph. Also fixed issues with adding units for fat, protein, etc

1.2. Jean Maldonado Caizapanta

- Added category labels to the hierarchical edge bundling diagram.
- Added color gradient to the lines in the hierarchical edge bundling diagram.
- Integrated hierarchical edge bundling diagram with initial layout,
- Fixed category image layout.
- Updated recipe layout with pie chart.
- Contributed to final report -- data section

1.3. Yang Qin

- Draw the sketch and layout of the potential design
- Wrote short descriptions and name for the project
- Completed the code for the donut chart
- Completed the code for the categories
- Added interactions between categories and donut chart, when the category is clicked, the recipes within donut chart get updated
- Contributed to final report -- Donut chart

2. Data

2.1. Source and structure

A recipes dataset from Kaggle (https://www.kaggle.com/hugodarwood/epirecipes) was used as a starting point for this project. This dataset contained a csv file with over 20,000 recipes from Epicurious (http://www.epicurious.com/). The dataset had over 600 variables which were predominantly ingredient names. There were also categories, states, holidays, and hashtags. All these variables had values either 0 or 1 to indicated whether the category had an ingredient, if it fell in a specific category or had a certain tag associated with it etc. The dataset also contained the following variables:

- Rating
- Calories
- Protein content
- Fat content
- Sodium content

After going through the dataset, we found some of these data was highly erroneous. We found over 200 recipes that had over 3000 calories. We ran searches for some recipes on the epicurious website and found that the calorie, protein, fat, and sodium values did not match the ones currently on the website. We also found that the following additional nutritional information was available:

- Carbohydrates
- Saturated fat content
- Fiber content
- Cholesterol content

We concluded that the dataset from Kaggle was out of date. Therefore, we scraped the epicurious website to retrieve all nutritional values using the current recipe names we had as search terms. This resulted in the new base data set which was used in the project.

2.2. Transformations and Integrations

It was necessary to both filter and transform the dataset in order to achieve the two visualizations that we envisioned to illustrate:

1. The relationships between ingredients.
2. Recipes with nutritional information.

We firstly cleaned the dataset to remove variables that were not useful for our visualizations. This included the removal of variables or states, hashtags, holidays, categories etc. The variables that ultimately remained were ingredient names and nutritional information.

Next, given that showing nutritional information was one of our goals, the first step was to filter out any recipes that did not have calorie information. We found that if calorie data was present, then nutritional information would accompany it.

In order to visualize relationships, a python script was developed to go through the original dataset and create a ‘relationships dataset’ where a record for every pair of ingredients

in the original dataset was generated. For example; if a recipe used egg, ham, and cheese as ingredients - then the script would produce a new record for each combination as follows:

- Egg --- Ham
- Egg --- Cheese
- Ham --- Cheese

The visualization we chose for showing ingredient relationships was Mike Bostock's hierarchical edge bundling (https://bl.ocks.org/mbostock/7607999). We decided to use the example code as a template and consequently faced technical challenges. Either change the code to parse our dataset or transform our dataset so we could reuse existing code. Transforming the dataset is much simpler so we transformed our 'relationships dataset' into a JSON 'hierarchical dataset'. The python script that was created went through the 'relationships dataset' and created a JSON object with an ingredient field and another field with a list of ingredients it was related to.

Considering that our dataset consisted of over 600 ingredients, we decided to split our data into the meal categories; Breakfast, Lunch, Brunch, Dinner, Dessert & Ice Cream, Vegetarian. Vegan, and Drink. Despite splitting the data into categories, some of these categories such as Dinner had a vast number of ingredient to ingredient relationships. Therefore, we also filtered the data by number of relationships. For example, for Dinner we filtered out relationships two ingredients were used together less than 15 times. We had different cutoffs for different meal categories since they all varied in size.

To create the colors of our dataset and group similar ingredients together on the graph, we went through every ingredient and assigned it to a category. We put these ingredient to category mappings in a new file which we then used as a lookup. We wanted to create enough categories so that you could see a clear distinction between different ingredients, but not enough so that each ingredient was in its own category. We settled on 14 different categories.

3. Building the Visualization

C. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales.

3.1. Hierarchical Edge Bundling

We used a color scale gradient that starts with the color of the path that corresponds to the color of the category of the source ingredient. The gradient end color would be the color of the category that corresponded to the ingredient it connects to. We accomplished this by segmenting the path into multiple different segments and assigning a different color to each of those segments as we found in another one of Mike Bostock's examples (https://bl.ocks.org/mbostock/4163057).

The position of the paths corresponded to where the ingredients were located in the diagram. We arranged the ingredients based on their categories, so each ingredient was in the same area as other ingredients of their categories.

3.2. Donut Chart

The donut chart is divided into two parts: dropdown menu and receipt composition chart

- Dropdown menus

Dropdown menus have: top rated recipes, top proteid recipes, less fat recipes, less HDL recipes, and less carbs recipes.

Top rated recipes: display the receipts ordered by ratings, the title is the name of the receipt.

Top proteid recipes: display the receipts ordered by protein-cotent

Less fat recipes: display the receipts desc ordered by fat-cotent

Less HDL recipes: display the receipts desc ordered by cholesterol-content

Less carbs recipes: display the receipts desc ordered by carbs-content

We display 5 recipes by default.

The default setting is "top rated recipes" in "breakfast" category.

- Receipt composition chart

The receipt composition chart shows the composition of calories, they are: protein, fat and carbs, according to the U.S. Department of Agriculture guidelines.

However, in our dataset, the protein, fat, carbs are measured by grams, thus in the donut chart, we need to convert the grams to calories, here is the formula:

Protein (g) * 4 = Protein (calories)

Carb (g) * 4 = Carb (calories)

Fat(g) * 9 = Fat (calories)

Total calories number = Protein (calories) + Carb (calories) + Fat (calories)

Take an example of the first row data:

Receipt itle: Lentil, Apple, and Turkey Wrap

protein-content: 31; fat-content: 10; carbs-content: 56

Then protein contains 31*4 = 124 calories, carb contains 56*4=224 calories, fat contains 10*9 = 90 calories, and total calories is: 124+224+90 = 438.

In the chart, we use 3 different colors to show fat, protein and carbs respectively, the legend can be seen as below:

4. Story

There were certain ingredients that were used with many other types of ingredients, regardless of the meal category. We found many relationships with eggs. This is not surprising

because eggs can be used in baking and for cooking as well. Other vegetables like onions and tomatoes also had many relationships with other ingredients in a variety of categories including herbs, meat, herbs, and condiments, and seafood. This suggests that onions and vegetables are versatile ingredients that could be added to many different types of ingredients.

There are certain ingredients that aren't common during some mealtimes. This includes seafood during breakfast, spices during brunch, and nuts during dinner. During breakfast, there are higher fat meats such as pork, sausage, and bacon, which are less common during dinner. The trend of having high fat recipes is common when looking at the highest rated recipes for the breakfast recipes as well, even when the selected ingredient is not a meat. For top rated breakfast recipes for cinnamon, cheese, and even peppers, many of these recipes contained recipes that are over 50% fat.

This higher fat trend continues during dinner as well. The top recipes for rosemary, tarragon, rosemary, or vegetables such as chile peppers, poblano peppers, and spinach have fat percentages that are higher than 50%. This is not to say that all recipes have high fat contents, but the highest rated have high fat content. These recipes also have high calorie counts as well, such as as the grilled lamb chops which are the highest rated dinner recipe for mushrooms. Cooking at home is generally seen as healthier than going to eat out at a restaurant, but we may need to pay more careful with making healthier choices at home as well.