# **Food choices and preferences of college students**

**Group 2:**  Yuki Yan, Yubing Yang, Yuxiang Wang, Zhe zhao

**Theme of Research Project**

In our daily life, dietary habits are becoming an important element that impacts a lot of aspects. As college students, our group interested in the food choices and preferences of college students. Examining the eating practices of college students is always relevant to improving health education. Today, a lot of college students have trouble keeping good dietary habits. Some of them stress eat, some eat food with little nutrition, some don’t have regular meals. This research project focuses on the food choices and preferences of college students, using the dataset to analyze the current dietary habits of students and nutrition information of college students. We also want to use ‘business logic’ to discover the internal relations for each attribute from the dataset.

Our dataset is a summary of 126 student‘s responses. The dataset includes information on food choices, GPA, nutrition, preferences on food type, parents' influence, the relativity of health and other data collected from the survey. This dataset contains 61 columns and various types including nutrition information, students’ taste preference of different cuisines, their parents cook, sports, weight, grade. Because our original dataset contains very extensive information,(like demographics, sports, parents, etc) not only for food, So we consider a comprehensive database to build upon because this dataset includes several different fields with different data. Building a database makes data management more easily. We can discover more relationships between fields once the data is structured and organized, collecting and storing more new data into the same database is also considerable and efficient. More specific, our group selected columns that we are interested in from 61 columns as a new dataset, classified by characteristics to eight fields, which are demographics, habits, parents, food-type, sports, calories, healthy and comfort\_food. They used a new dataset to analyze the relations which are inside each field or relations between each field. Then we listed the questions which we are interested in, like What’s the favorite\_type for college students? What’s the relation between GPA and veggies\_day? What the difference between male and female thinking about calories? How important is nutrition information for today's college students? Are these kids likely to have a different taste compared to others? etc.

**Tables, relationships, and/or ‘themes’ that appear in database**

To analyze the food choices and preferences of college students comprehensively, We build a mean and important table called the demographic which is the basic information about each student, setting the id as the primary key and we create 7 subtables. The other words to effect on our project in our data are the weight which has a stronger relationship with the eating practices of college students. Therefore, we want to see how the weight change by how many calories in taking (calorie table), what kind of sports is doing in free time (sports table) and what kind of food is taking (food\_type table). Moreover, there exists a phenomenon among college students. Most college students prefer eating outside or calling a delivery to cook themselves or parent cooking. So, parents table, eating habits table, and comfort\_food table can help us to figure out this phenomenon by associating the keywords (income, pay\_meal\_out, and on\_off\_campus) in the demographics table.

The last question we care about is the health versus food choices of college students. Since college students seem to be fashion, diverse thinking, and activity, the food choices are in many fields among the students. Therefore whether food healthy or not is the key terms we are concerned with. To solve this issue, we can merge all the information we have, not only focus on the habits of students but also need to check the health value of the food which college students choice. Since many factors will affect the changes in health. We create a big table(health table) that contains the type of cuisine, nutrition of the food, and the vitamins. After combining the table (eating habits), we can clarify the healthy of the situation for the current eating habits of college students.

For business logic, we can use that information which combines with different tables to predict our potential customers if we want to join the food market. For example, after combining the demographic table and food\_type table, we can find a result that most people who like Indian food weight 180IB, also from these two tables, we can see that most people whose GPA is greater than 3.5 preferring to eating Chinese food. After analysis of our data, we can pretty much clarify the food choice and preferences of college students, then we can suggest food court, schools' dining hall or other food companies improve or change the health food type which most students are willing to eat. Besides, if students want to increase their GPA, what kind of food will help them to improve.

**Relevant/Important Columns/Fields/Data**

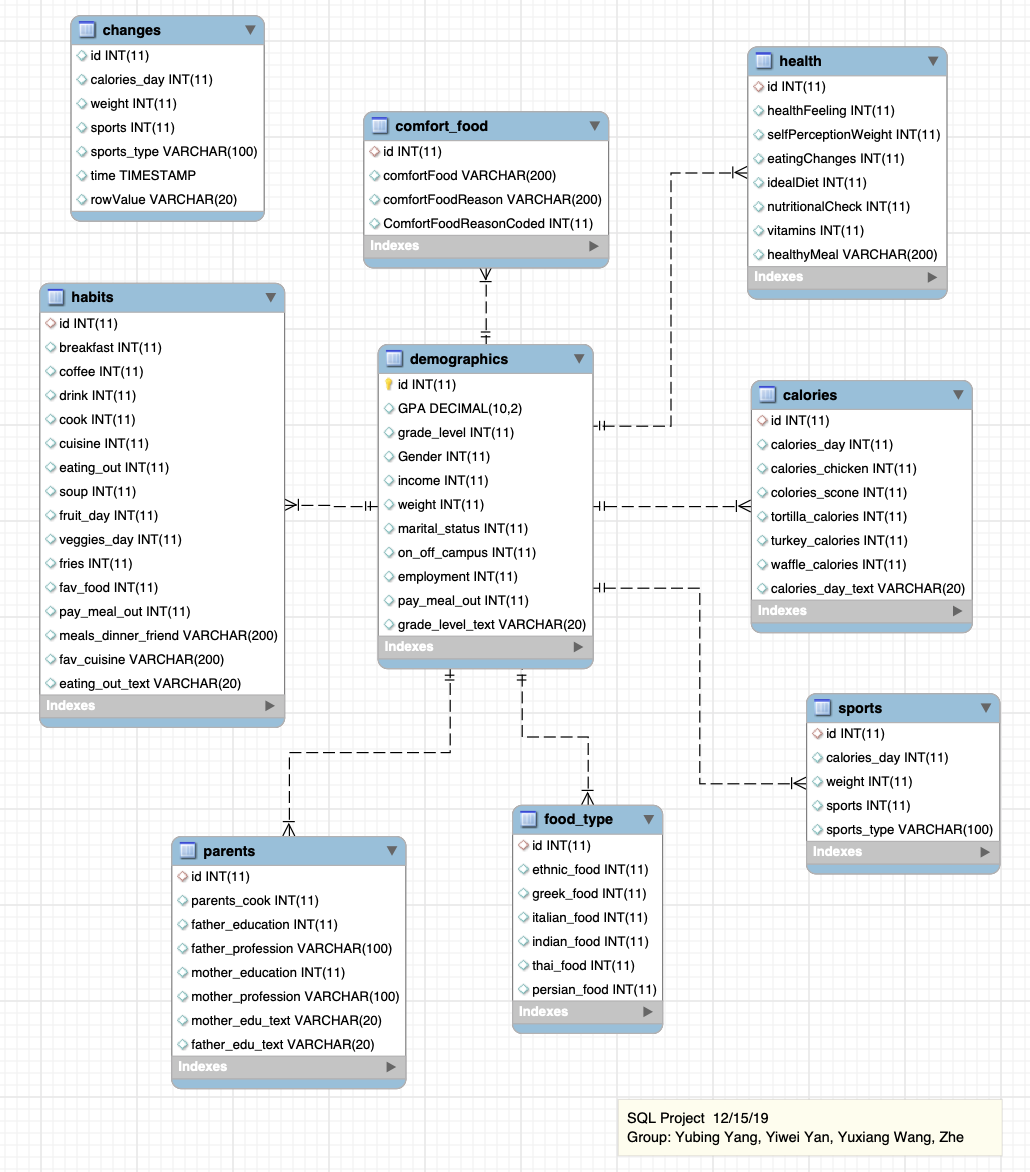
Our group found our original dataset, it’s an uncleared dataset, so we did data cleaning as our first step. We found there are some missing values and NA values. We changed them to 0 because our original dataset is a summary of responses in which most columns are numeric type (means rating). Thus, there are no influences and even better to analyze and insert data to tables. We only use this one dataset with plenty of columns for the project. Since this is a survey dataset asking questions about the food choices of college students, we do not have proper any existing column as the primary key. To build connections and relationships within the dataset, also for further analysis, we create a column called id with type as an integer in the demographic column and add foreign key for each other table referencing it.

**Transformations**

The original dataset has 61 columns and contains very extensive information. Our group found some columns can be regarded as one field. Therefore, we selected columns that we are interested in and clustered them to 8 fields as 8 tables as our new dataset (demographics, habits, parents, food-type, sports, calories, healthy and comfort\_food). Demographics table describes the demographics information for each student who did the survey (like, GPA, gender, etc). Some columns in this table are very important ( like gender, GPA, grade\_level, on\_off\_campus) to group by students then related to other tables to find relations and differences. Habits table describes the eating habits for college students which can help us to analyze the performance and lifestyle of eating. Parents' table describes the information of student’s parents then helps us to analyze the impacts of parents for their dietary habits. Food\_type table describes the rating number of each\_food Which can help us to know the most popular food for college students. The calorie table describes the calories for different kinds of foods. The healthy table describes the self-healthy evaluation and the comfort\_food table shows the rating of comfort\_food. The demographics table is the most important one which is relating to other tables. Our group selected the main columns from each table to get the correlation, then analyze the Food choices and preferences of college students.

While writing the select statements, we found out there is some relevant column with not text explanations. To let the statements make more sense, we altered several tables to add explaining columns for some numbered columns such as ComfortFoodReasonCoded, grade\_leve, eating\_out, and calories\_day, then add values by updating. This way helps our group to analyze the result more directly and easily.

**ER Diagram**

****

**Select statements**

* 1.1 How many people for each comfort food reason?

We got most students choosing comfort food because of boredom.

Asking this question helps to know the most popular reason for college students buying comfort food, so that food brand or retailer can use this information to build products or advertisements.

select comfortFoodReasonCoded, comfortFoodReason, count(\*) as numOfStudents from comfort\_food

group by comfortFoodReasonCoded,comfortFoodReason order by numOfStudents desc;

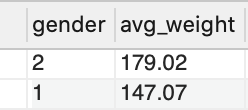


* 1.2 What are the average weights for each gender (1:male, 2:female)?

We Get an average weight for males who are 179.02 and females are 147.07.

In this statement, we want to know the average weight for each gender. From this result, we can see both the mean of male and female are in a reasonable area, which helps us to clear our data are reasonable, without lots of outliers. It also can help us a stronger conclusion for another question.

select gender, round(avg(weight),2) as avg\_weight from demographics where weight <> 0 group by gender;

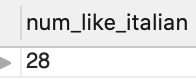


* 1.3 How many students like Italian food?

We find Italian food is the most favorite food among the students.

This statement can help us to know the most popular cuisine, from the economics’ perspective, the Italian restaurant might gain more profit around the college. By selecting each type of food, we can find that people who choose Italian food is the most, the total is 28. After combining other food types, we can make a conclusion which is Italian is popular among the students.

select count(id) as num\_like\_italian from habits where fav\_cuisine like'Italian';



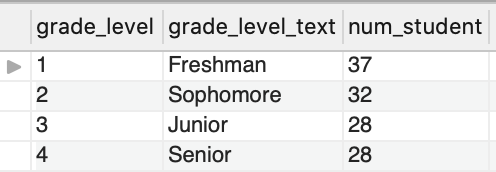
* 1.4 How many students for each grade level?

We found most students are Freshmen who did this survey.

This question can help us to know the grade\_level we should focus on If we want to do some changes. Since the different grades may because of the change of the food choice, therefore, we are doing specific demography to check the range in different grades, from the result, we find the standard deviation for each grade is not very huge. Therefore, It seems that our data do not have a bias on the grade chosen.

select grade\_level, grade\_level\_text, count(\*) as num\_student from demographics

group by grade\_level, grade\_level\_text order by grade\_level;



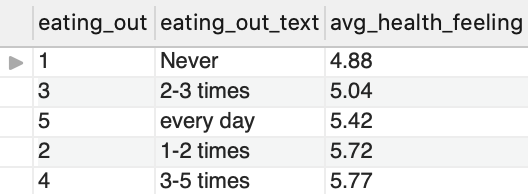
* 2.1 Show the average health feeling for each level of eating out frequency.

We found students who eating out 3-5 times/week have the highest score (max 10) and students never eating out have the lowest score.

Because eating out is a popular phenomenon in our daily life, and we can’t guarantee each meal is healthy. So this question helps us to know the relationship between students’ self-health perception and frequency of eating out. We can see that eating out less frequently makes students more likely to think of themselves very healthy.

select eating\_out, eating\_out\_text, round(avg(healthFeeling),2) as avg\_health\_feeling from habits join health using(id)

group by eating\_out, eating\_out\_text order by avg\_health\_feeling;



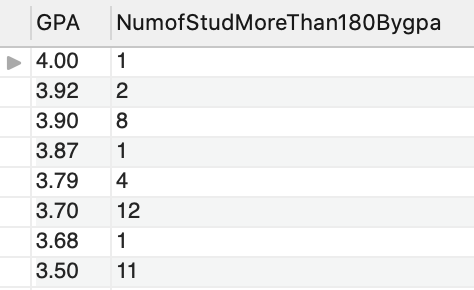
* 2.2 Show the number of students who weigh more than 180 lb in terms of GPA from 3.5.?

We found most students got a 3.7 GPA if they over 180 lbs.

We chose this statement because we try to find the relation between GPA and student’s weight.180 is a normal weight for males.

select GPA, count(\*) as NumofStudMoreThan180Bygpa from demographics left join sports using(weight)

where weight > 180 and GPA >= 3.5 and GPA <> 0 group by GPA order by GPA desc;



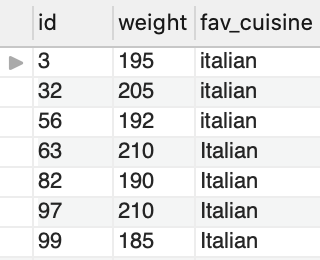
* 2.3 Show students whose weight are more than 180lb like Italian food.

We found the range of weight is 185 - 210.

We chose this statement because we want to know the weight range of students who prefer Italian food and weight more than 180lbs.

select id,weight,fav\_cuisine from habits join demographics as d using(id)

where weight>=180 and fav\_cuisine like 'Italian';



* 2.4 How many males and females who think consuming calories per day is very important(1:Female, 2:male)

We found 13 Females and 10 Males think calories is very important. Around 17%.

Since the current college students have poor knowledge on the calories, we want to see how many boys and girls frequently care about their calorie intake, which is important for us to make a right analyze on the nutrition table.

select d.Gender,count(\*) as num\_students,c.calories\_day, calories\_day\_text from demographics d inner join calories c using(id)

where calories\_day = '4' group by d.Gender, calories\_day, calories\_day\_text order by gender;

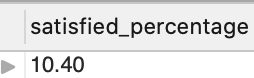


* 3.1 what is the percentage of students who are pretty satisfied with their current diets?

Seeing only 10.4% of college students are satisfied with their current diet. We want to know the current situation of students about their current diets. The college could think of how to improve food on campus since most students eat on campus more or less.

select round(currentdiet\*100/total,2) as satisfied\_percentage from(

select (select count(\*) from health where idealDiet = 6) currentdiet,count(\*) as total from health) sub;



* 3.2 what is the percentage of students with grade higher than 3.5 very likely to eat veggies and fruits every day?

We found 37.5% of students eat veggies and fruits every day then they got a pretty good GPA.

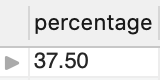
We chose this statement because want to know if students who take the dietary fiber and vitamins daily from veggies and fruits do better than others or not.

select round(goodgrades\_vf \* 100 / goodgrades\_tt,2) as percentage from(

select (select count(\*) from habits where veggies\_day =5 and fruit\_day =5 and

id in (select id from demographics where gpa >= 3.5)) goodgrades\_vf, count(\*) as goodgrades\_tt from demographics where gpa >=3.5) sub ;

select count(\*) from demographics where gpa >= 3.5;



* 3.3 count how many students with income higher than 100,000 do not think themselves very healthy.

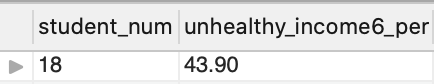
The result is surprised us because there are 18 students who get an income higher than 100000 and 44% of them think they are unhealthy.

Since most students are raised by their families, we want to determine the reflection about the current health level from those families who get high pay per year. Then we can make a deeper analysis of the food type choice and performance.

select student\_num, round(student\_num \*100/ (select count(\*) from demographics where income > 5),2) unhealthy\_income6\_per

from( select count(\*) as student\_num from demographics

where income > 5 and id in (select id from health where healthfeeling <=5)) income6unhealthy;

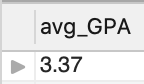


* 3.4 What is the average GPA for students who are likely to eat veggies in a day?

We want to see is there a relationship between eating habits and GPA. Thus, we try to only use one variable to check the change of GPA. Although we know only use one variable is not enough, we just want to see is a big effect only by changing one variable. And we find the people frequently eat veggies are getting a higher GPA which is 3.37.

select round(avg(GPA),2) as avg\_GPA FROM demographics

where id in (select id FROM habits where veggies\_day > 3);



**Reference**

Pajo,B. (2016). Food choices: College students' food and cooking preferences. *Kaggle.com.* Retrieved from: <https://www.kaggle.com/borapajo/food-choices>