How hazardous are college students' food preferences to their well being?

Head Detective: Brenda Xu
Email: bx320@stern.nyu.edu

For my project, I will be using two data sets I found on Kaggle, in order to see if the food preferences and eating habits of college students put them at a greater risk of certain adverse food events than the average American. The two datasets are: Adverse Food Events (https://www.kaggle.com/fda/adverse-food-events) and Food Choices of College Students (https://www.kaggle.com/borapajo/food-choices).

Access: All of the data I will use can be downloaded off the website as a zip file. For the food choices file, the zip file contains an Excel sheet with all the data, as well as a Word document that gives some supplementary information about notations. For example, survey participants put a "1" in the Gender column if they were female, and a "2" if they were male. For those columns, I will eventually convert the numbers to data that is easily understandable without an extra document.

The adverse food events file contains an Excel file with the data and a PDF with explanations of the data.

All the data will be accessed through the read_csv() function. For simplicity, I placed both Excel documents into my current folder before starting.

In [21]:

```python
#below are the required packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
In [22]: adverse = pd.read_csv("CAERS_ASCII_2004_2017Q2.csv") #this is the name of the adv
         adverse.head(5)
```

Out[22]:

	RA_Report #	RA_CAERS Created Date	AEC_Event Start Date	PRI_Product Role	PRI_Reported Brand/Product Name	PRI_FDA Industry Code	PRI_FDA Indust Nam
0	65325	1/1/2004	8/4/2003	Suspect	MIDWEST COUNTRY FAIR CHOCOLATE FLAVORED CHIPS	3	Bake Prod/Dough/Mix/Icir
1	65325	1/1/2004	8/4/2003	Suspect	MIDWEST COUNTRY FAIR CHOCOLATE FLAVORED CHIPS	3	Bake Prod/Dough/Mix/Icir
					KROGER CLASSIC		

It looks like most or all of those categories are useful, so I will leave all the columns in but rename them.

```
In [25]: adverse.columns = ["Report ID", "Event Entered", "Event Start Date", "Suspect or
                            "Industry Name", "Consumer Age", "Units of Age", "Gender", "Out
```

Below is adverse, with columns renamed.

```
In [2]: adverse.head(5)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-e797ba01ef75> in <module>()
----> 1 adverse.head(5)

NameError: name 'adverse' is not defined
```

```
In [27]: college_food = pd.read_csv("food_coded.csv") #reading in the data about college s
```

In [28]:
```
college_food.head(5)
```

Out[28]:

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food
0	2.4	2	1	430	NaN	315.0	1	none
1	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream
2	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food
3	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream
4	3.5	1	1	720	2.0	420.0	2	Ice cream, chocolate, chips

5 rows × 61 columns

There are several columns that I do not plan on using, such as "tortilla_calories" in college_food. This column asked survey participants how many calories they thought were in a burrito from Chipotle. I'll only take the columns I think will be useful.

In [29]:
```
college_food1 = college_food[["Gender","comfort_food", "food_childhood", "fruit_d
college_food1.head()
#I may change what columns I include if I feel like they may add to my analysis. 
#a scale of 1-5, how likely they are to try Indian food.

#I included vitamins because in the adverse food events file, there is an option 
```

Out[29]:

	Gender	comfort_food	food_childhood	fruit_day	meals_dinner_friend	vitamins
0	2	none	rice and chicken	5	rice, chicken, soup	1
1	1	chocolate, chips, ice cream	chicken and biscuits, beef soup, baked beans	4	Pasta, steak, chicken	2
2	1	frozen yogurt, pizza, fast food	mac and cheese, pizza, tacos	5	chicken and rice with veggies, pasta, some kin...	1
3	1	Pizza, Mac and cheese, ice cream	Beef stroganoff, tacos, pizza	4	Grilled chicken \rStuffed Shells\rHomemade Chili	1
4	1	Ice cream, chocolate, chips	Pasta, chicken tender, pizza	4	Chicken Parmesan, Pulled Pork, Spaghetti and m...	2

I plan to come up with a way to rank the severity of symptoms that are in adverse["Symptoms"]. Since there are a lot of different symptoms, I might come up with an arbitrary way to rank them, such as nausea being 0 for least severe, and death being 100 for most severe. Another way to control for the huge number and variation in data overall, is to focus on just a few symptoms, for example nausea, death, rash, vomiting.

If I go with the second option, I want to select a symptom by using .match("symptom") and making a new dataframe using only instances where that symptom showed up. For example, I would create a dataframe with only rows that had "nausea" listed as a symptom. There would be dataframes for each symptom I choose.

I would then go through the food_choices dataframe, and match foods that college students like, to the ["Industry Name"] or ["Product Name"] columns under the adverse dataframe. For example, a lot of students seem to like ice cream as their comfort food. That is easily matched to "Ice cream prod" entries under ["Industry Name"]. In each of my symptom dataframes, I would then either use a count or percentage to check the severity of rate of occurence. My conclusions might sound something like: "Ice cream appears in x% of entries under "nausea" but 0 times under "death". Fruit appears many times under death, and a lot of college students say they eat fruit frequently, so this fruit-eating habit is more hazardous to their well-being than their ice cream habit."

```
In [1]: deaths = adverse[adverse["Symptoms"].dropna().str.match('DEATH')]
        deaths.head()
        #Question: I tried to make the symptoms dataframes that
        #I mentioned above but I keep getting an error. How do I solve this?
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-3a35f20f3845> in <module>()
----> 1 deaths = adverse[adverse["Symptoms"].dropna().str.match('DEATH')]
      2 deaths.head()
      3 #Question: I tried to make the symptoms dataframes that I mentioned abo
ve but I keep getting an error. How do I solve this?

NameError: name 'adverse' is not defined
```

Summary

I have the data I need and I'm confident that I can extract and use it, but I can't get past the IndexingError above.

```
In [ ]:
```