Module 1 Exercises for Python

Filipp Krasovsky, 3-8-21

In [2]:
```
import pandas as pd
import numpy as np
```

In [5]:
```
#import data set
df = pd.read_csv("C:/Users/Filipp/Documents/usd_data_sci/502_data mining/module1/Website Data Sets/nutrition_subset.csv")
```

In [8]:
```
#sanity check
df.head()
```

Out[8]:

	food item	weight_in_grams	saturated_fat	cholesterol
0	GELATIN; DRY 1 ENVELP	7.00	0.0	0
1	SEAWEED; SPIRULINA; DRIED 1 OZ	28.35	0.8	0
2	YEAST; BAKERS; DRY; ACTIVE 1 PKG	7.00	0.0	0
3	PARMESAN CHEESE; GRATED 1 OZ	28.35	5.4	22
4	PARMESAN CHEESE; GRATED 1 CUP	100.00	19.1	79

Question 21 - Ch.3 The elements in the data set are food items of various sizes, ranging from a teaspoon of cinnamon to an entire carrot cake. a. Sort the data set by the saturated fat (saturated_ fat) and produce a listing of the five food items highest in saturated fat. b. Comment on the validity of comparing

In [12]:
```
df_sorted = df.sort_values(by=['saturated_fat'],ascending=False)
df_sorted.head()
```

Out[12]:

	food item	weight_in_grams	saturated_fat	cholesterol
378	CHEESECAKE 1 CAKE	1110.0	119.9	2053
535	ICE CREAM; VANLLA; RICH 16% FT1/2 GAL	1188.0	118.3	703
458	YELLOWCAKE W/ CHOCFRSTNG;COMML1 CAKE	1108.0	92.0	609
581	CREME PIE 1 PIE	910.0	90.1	46
890	LARD 1 CUP	205.0	80.4	195

Comparrisons are not valid due to the variation in serving sizes (weight_in_grams).

Question 22 - Ch. 3 Derive a new variable, saturated_ fat_per*gram, by dividing the amount of saturated fat by the weight in grams. a. Sort the data set by saturated* fat_per_gram and produce a listing of the five food items highest in saturated fat per gram. b. Which food has the most saturated fat per gram?

In [16]:
```
df['saturated_fat_per_gram'] = df['saturated_fat'] / df['weight_in_grams']
df_sorted = df.sort_values(by=['saturated_fat_per_gram'],ascending=False)
df_sorted.head()
```

Out[16]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram
908	BUTTER; SALTED 1 TBSP	14.0	7.1	31	0.507143
909	BUTTER; UNSALTED 1 TBSP	14.0	7.1	31	0.507143
710	BUTTER; UNSALTED 1/2 CUP	113.0	57.1	247	0.505310
709	BUTTER; SALTED 1/2 CUP	113.0	57.1	247	0.505310
913	BUTTER; UNSALTED 1 PAT	5.0	2.5	11	0.500000

The food with the highest saturated fat per gram is butter (salted)

Question 23 - Ch. 3

Derive a new variable, cholesterol_per_gram. a. Sort the data set by cholesterol_per_gram and produce a listing of the five food items highest in cholesterol fat per gram. b. Which food has the most cholesterol fat per gram?

In [18]:
```
df['cholesterol_per_gram'] = df['cholesterol']/df['weight_in_grams']
df_sorted = df.sort_values(by=['cholesterol_per_gram'],ascending=False)
df_sorted.head()
```

Out[18]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol_per_gram
119	EGGS; RAW; YOLK 1 YOLK	17.0	1.6	213	0.094118	12.529412
58	CHICKEN LIVER; COOKED 1 LIVER	20.0	0.4	126	0.020000	6.300000
45	BEEF LIVER; FRIED 3 OZ	85.0	2.5	410	0.029412	4.823529
167	EGGS; COOKED; FRIED 1 EGG	46.0	1.9	211	0.041304	4.586957
186	EGGS; COOKED; HARD-COOKED 1 EGG	50.0	1.6	213	0.032000	4.260000

The food with the most cholesterol per gram is eggs.

Question 24 - ch.3

Standardize the field saturated_fat_per_gram. Produce a listing of all the food items that are outliers at the high end of the scale. How many food items are outliers at the low end of the scale?

In [30]:
```
from scipy import stats

df['sfpg_z'] = stats.zscore(df['saturated_fat_per_gram'])
#sanity check
pos_outliers = df.query('sfpg_z > 3')
neg_outliers = df.query('sfpg_z < -3')

print(pos_outliers[['food item','sfpg_z']])
```

```
                                food item    sfpg_z
210  CHOCOLATE; BITTER OT BAKING   1 OZ      4.240676
448  COCONUT; RAW; SHREDDED        1 CUP     3.938687
492  COCONUT; DRIED; SWEETND;SHREDD1 CUP     4.204266
576  COCONUT; RAW; PIECE           1 PIECE   3.942889
709  BUTTER; SALTED                1/2 CUP   7.082741
710  BUTTER; UNSALTED              1/2 CUP   7.082741
890  LARD                          1 CUP     5.371375
898  FATS; COOKING/VEGETBL SHORTENG1 TBSP    3.278227
899  LARD                          1 TBSP    5.373078
907  FATS; COOKING/VEGETBL SHORTENG1 CUP     3.223726
908  BUTTER; SALTED                1 TBSP    7.110475
909  BUTTER; UNSALTED              1 TBSP    7.110475
912  BUTTER; SALTED                1 PAT     7.002408
913  BUTTER; UNSALTED              1 PAT     7.002408
920  IMITATION CREAMERS; POWDERED  1 TSP     4.732985
```

In [31]:
```
print(len(neg_outliers))
```

```
0
```

there are no outliers left of the measure of central tendency.

Question 25 - ch. 3 Standardize the field cholesterol_per_gram. Produce a listing of all the food items that are outliers at the high end of the scale.

In [33]:
```
df['cpg_z'] = stats.zscore(df['cholesterol_per_gram'])
pos_outliers = df.query('cpg_z > 3')
print(pos_outliers[['food item','cpg_z']])
```

```
                               food item      cpg_z
45   BEEF LIVER; FRIED             3 OZ     6.765448
58   CHICKEN LIVER; COOKED         1 LIVER  8.952391
119  EGGS; RAW; YOLK               1 YOLK  18.179372
167  EGGS; COOKED; FRIED           1 EGG    6.415037
184  EGGS; RAW; WHOLE              1 EGG    5.930750
185  EGGS; COOKED; POACHED         1 EGG    5.901127
186  EGGS; COOKED; HARD-COOKED     1 EGG    5.930750
189  EGGS; COOKED; SCRAMBLED/OMELET1 EGG    4.841464
```