Rudy Fasano

Applied Data Mining

Module 1

Chapter 3 - # 21-26

In [9]:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
%matplotlib inline
```

In [2]:

```
df = pd.read_csv("D:/School/502/Week1/datasets/Website Data
Sets/nutrition_subset.csv")
df.head()
```

Out[2]:

	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 - Sorting saturated fat and commenting on validity

In [3]:

```
### q21: a) sort the dataset and list top 5 in highest saturated fat
sort_by_satfat = df.sort_values('saturated_fat', ascending=False)
sort_by_satfat.head()
```

Out[3]:

	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

	food item	weight_in_grams	saturated_fat	cholesterol
890	LARD 1 CUP	205.0	80.4	195

Question 22 - Derive a new variable, saturated_fat_per_gram, by dividing the amount of saturated fat by the weight in grams

In [6]:

```
### q22: a) create new variable and list the five food items highest in saturated fat per gram
df['saturated_fat_per_gram'] = df['saturated_fat']/df['weight_in_grams']
## q22: b) sort in descending order to find most saturated fat per gram
sat_fat_sorted = df.sort_values('saturated_fat_per_gram', ascending=False)
sat_fat_sorted.head()
```

Out[6]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol_per_gra
908	BUTTER; SALTED 1 TBSP	14.0	7.1	31	0.507143	2.21428
909	BUTTER; UNSALTED 1 TBSP	14.0	7.1	31	0.507143	2.21428
710	BUTTER; UNSALTED 1/2 CUP	113.0	57.1	247	0.505310	2.18584
709	BUTTER; SALTED 1/2 CUP	113.0	57.1	247	0.505310	2.18584
913	BUTTER; UNSALTED 1 PAT	5.0	2.5	11	0.500000	2.20000

Question 22 - b) which food has the most saturated fat per gram? - Both 1 tablespoon of salted butter and 1 tablespoon of unsalted butter both have the most saturated fat per gram (0.507)

Question 23 - Derive a new variable, cholesterol_per_gram

In [8]:

```
### Q23: a) create new variable and sort to find five food items highest in cholesteral fat per gram
df['cholesterol_per_gram'] = df['cholesterol']/df['weight_in_grams']

### sorting in descending order then identifying top 5 food items in cholesterol per gram
satfat_sort = df.sort_values('cholesterol_per_gram', ascending=False)
satfat_sort.head()
```

Out[8]:

	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

Question 23 - b) Which food has the most cholesterol fat per gram? - 1 raw egg yolk is listed to have the most cholesterol fat per gram (12.53).

Question 24 - Standardize the field saturated_fat_per_gram. Produce a listing of all the food items that are outliers at the end of the scale. how many food items are outliers at the low end of the scale?

In [15]:

```
### standardize saturated_fat_per_gram
df['saturated_fat_per_gram_z'] = stats.zscore(df['saturated_fat_per_gram'])
df.query('saturated_fat_per_gram_z > 3 | saturated_fat_per_gram_z < -3')
### identifying outliers
```

Out[15]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol
210	CHOCOLATE; BITTER OT BAKING 1 OZ	28.35	9.0	0	0.317460	
448	COCONUT; RAW; SHREDDED 1 CUP	80.00	23.8	0	0.297500	
492	COCONUT; DRIED; SWEETND;SHREDD1 CUP	93.00	29.3	0	0.315054	

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol
576	COCONUT; RAW; PIECE 1 PIECE	45.00	13.4	0	0.297778	
709	BUTTER; SALTED 1/2 CUP	113.00	57.1	247	0.505310	
710	BUTTER; UNSALTED 1/2 CUP	113.00	57.1	247	0.505310	
890	LARD 1 CUP	205.00	80.4	195	0.392195	
898	FATS; COOKING/VEGETBL SHORTENG1 TBSP	13.00	3.3	0	0.253846	
899	LARD 1 TBSP	13.00	5.1	12	0.392308	
907	FATS; COOKING/VEGETBL SHORTENG1 CUP	205.00	51.3	0	0.250244	
908	BUTTER; SALTED 1 TBSP	14.00	7.1	31	0.507143	
909	BUTTER; UNSALTED 1 TBSP	14.00	7.1	31	0.507143	
912	BUTTER; SALTED 1 PAT	5.00	2.5	11	0.500000	
913	BUTTER; UNSALTED 1 PAT	5.00	2.5	11	0.500000	
920	IMITATION CREAMERS; POWDERED 1 TSP	2.00	0.7	0	0.350000	

Question 24 cont. , there are no outliers below the lower limits, only above the upper limits as displayed above with a total of 15.

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

In [16]:

```
### standardize saturated_fat_per_gram
df['cholesterol_per_gram_z'] = stats.zscore(df['cholesterol_per_gram'])
df.query('cholesterol_per_gram_z > 3 | cholesterol_per_gram_z < -3') ###
identifying outliers
```

Out[16]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	choleste
45	BEEF LIVER; FRIED 3 OZ	85.0	2.5	410	0.029412	
58	CHICKEN LIVER; COOKED 1 LIVER	20.0	0.4	126	0.020000	

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	choleste
119	EGGS; RAW; YOLK 1 YOLK	17.0	1.6	213	0.094118	
167	EGGS; COOKED; FRIED 1 EGG	46.0	1.9	211	0.041304	
184	EGGS; RAW; WHOLE 1 EGG	50.0	1.6	213	0.032000	
185	EGGS; COOKED; POACHED 1 EGG	50.0	1.5	212	0.030000	
186	EGGS; COOKED; HARD-COOKED 1 EGG	50.0	1.6	213	0.032000	
189	EGGS; COOKED; SCRAMBLED/OMELET1 EGG	61.0	2.2	215	0.036066	

Question 25 cont. The identified outliers above the upper limits as displayed above with a total of 8.

Question 26 - Add a record index field to the data set.

In [49]:

```
## adding a record index field to data set
df2.shape


df2['index'] = pd.Series(range(0,961))  #adding index
df2.head()
```

Out[49]:

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol_per_gram
0	GELATIN; DRY 1 ENVELP	7.00	0.0	0	0.000000	0.000000
1	SEAWEED; SPIRULINA; DRIED 1 OZ	28.35	0.8	0	0.028219	0.000000
2	YEAST; BAKERS; DRY; ACTIVE 1 PKG	7.00	0.0	0	0.000000	0.000000
3	PARMESAN CHEESE; GRATED 1 OZ	28.35	5.4	22	0.190476	0.776014

	food item	weight_in_grams	saturated_fat	cholesterol	saturated_fat_per_gram	cholesterol_per_gram
4	PARMESAN CHEESE; GRATED 1 CUP	100.00	19.1	79	0.191000	0.790000