Module 1 Homework

Filipp Krasovsky

3/8/2021

Assignment: Data Science Using Python and R: Chapter 3 - Page 45: Questions #21, 22, 23, 24, & 25 Hint: Use both R and Python for these questions. Hint: Datasets for this assignment are available on the Weekly Python & R with Datasets page or you can download them here.

For Exercises 21–25, work with the Nutrition_subset data set. The data set contains the weight in grams along with the amount of saturated fat and the amount of cholesterol for a set of 961 foods. Use either Python or R to solve each problem.

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

```
#Chapter 3

#read our nutritional data in
df = read.csv("C:/Users/Filipp/Documents/usd_data_sci/502_data mining/module1/Website Data Sets/nutriti
```

```
#QUESTION 21

#a:sort by saturated fat
df.sorted_desc = df[order(df$saturated_fat,decreasing=TRUE),]

#get top five food items
df.top_five = head(df.sorted_desc,n=5)

#print our top five
print(df.top_five)
```

```
##                                 food.item weight_in_grams saturated_fat
## 379 CHEESECAKE                   1 CAKE              1110         119.9
## 536 ICE CREAM; VANLLA; RICH 16% FT1/2 GAL            1188         118.3
## 459 YELLOWCAKE W/ CHOCFRSTNG;COMML1 CAKE             1108          92.0
## 582 CREME PIE                    1 PIE               910          90.1
## 891 LARD                         1 CUP               205          80.4
##     cholesterol
## 379        2053
## 536         703
## 459         609
## 582          46
## 891         195
```

b. the comparison of saturated fat is not meaningful because we are looking at different serving sizes, even across the same food group (ex: Parmesan Cheese).

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?

```
#create our new variable
df$saturated_fat_per_gram = df$saturated_fat/df$weight_in_grams
#a. sort by saturated fat per gram
df.sorted_by_sfpg = df[order(df$saturated_fat_per_gram,decreasing=TRUE),]
#b. get the food with the most saturated fat per gram
most_sfpg = subset(head(df.sorted_by_sfpg,n=1),select=c(food.item,saturated_fat_per_gram))

print(most_sfpg)
```

```
##                                 food.item saturated_fat_per_gram
## 909 BUTTER; SALTED               1 TBSP                0.5071429
```

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?

```
#create new variable
df$cholesterol_per_gram = df$cholesterol/df$weight_in_grams

#sort dataset, produce top five, output food with the most cholesterol
df.sorted_chol = df[order(df$cholesterol_per_gram,decreasing=TRUE),]
df.top_five_chol = subset(head(df.sorted_chol,n=5),select=c(food.item,cholesterol_per_gram))

print(df.top_five_chol)
```

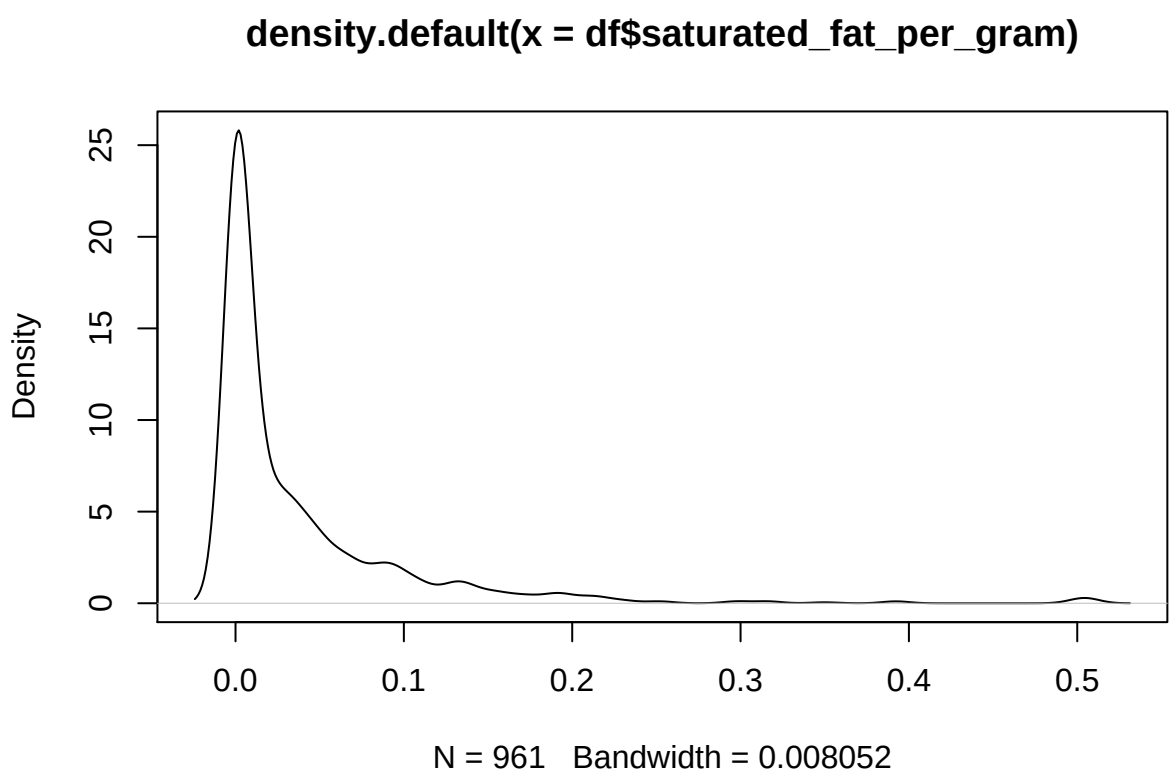
```
##                                 food.item cholesterol_per_gram
## 120 EGGS; RAW; YOLK              1 YOLK              12.529412
## 59  CHICKEN LIVER; COOKED        1 LIVER              6.300000
## 46  BEEF LIVER; FRIED            3 OZ                 4.823529
## 168 EGGS; COOKED; FRIED          1 EGG                4.586957
## 185 EGGS; RAW; WHOLE             1 EGG                4.260000
```

The food item 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?

```
#begin by looking at distribution
d <- density(df$saturated_fat_per_gram) # returns the density data
plot(d)
```



Our plot suggests that saturated_fat_per_gram does not follow a standard normal distribution, but rather a chi-square distribution of k~3. There are no outliers at the low end of the scale. Based on the proceeding chi-square test for outliers, we can conclude that we have at least one outlier on the high end. There are no outliers on the low end.

```
require(outliers)
```

```
## Loading required package: outliers
```

```
chisq.out.test(df$saturated_fat_per_gram)
```

```
##
##  chi-squared test for outlier
##
## data:  df$saturated_fat_per_gram
## X-squared = 50.506, p-value = 1.188e-12
## alternative hypothesis: highest value 0.507142857142857 is an outlier
```

```
#standardize
df$sfpg.std = scale(x = df$saturated_fat_per_gram)
subset(df,abs(df$sfpg.std) >= 3)
```

```
##                                food.item weight_in_grams saturated_fat
## 211 CHOCOLATE; BITTER OT BAKING   1 OZ               28.35           9.0
## 449 COCONUT; RAW; SHREDDED        1 CUP              80.00          23.8
```

```
## 493 COCONUT; DRIED; SWEETND;SHREDD1 CUP              93.00          29.3
## 577 COCONUT; RAW; PIECE           1 PIECE            45.00          13.4
## 710 BUTTER; SALTED                1/2 CUP           113.00          57.1
## 711 BUTTER; UNSALTED              1/2 CUP           113.00          57.1
## 891 LARD                          1 CUP             205.00          80.4
## 899 FATS; COOKING/VEGETBL SHORTENG1 TBSP             13.00           3.3
## 900 LARD                          1 TBSP             13.00           5.1
## 908 FATS; COOKING/VEGETBL SHORTENG1 CUP             205.00          51.3
## 909 BUTTER; SALTED                1 TBSP             14.00           7.1
## 910 BUTTER; UNSALTED              1 TBSP             14.00           7.1
## 913 BUTTER; SALTED                1 PAT               5.00           2.5
## 914 BUTTER; UNSALTED              1 PAT               5.00           2.5
## 921 IMITATION CREAMERS; POWDERED  1 TSP               2.00           0.7
##     cholesterol saturated_fat_per_gram cholesterol_per_gram sfpg.std
## 211           0              0.3174603            0.0000000 4.238469
## 449           0              0.2975000            0.0000000 3.936637
## 493           0              0.3150538            0.0000000 4.202078
## 577           0              0.2977778            0.0000000 3.940837
## 710         247              0.5053097            2.1858407 7.079055
## 711         247              0.5053097            2.1858407 7.079055
## 891         195              0.3921951            0.9512195 5.368580
## 899           0              0.2538462            0.0000000 3.276520
## 900          12              0.3923077            0.9230769 5.370282
## 908           0              0.2502439            0.0000000 3.222049
## 909          31              0.5071429            2.2142857 7.106775
## 910          31              0.5071429            2.2142857 7.106775
## 913          11              0.5000000            2.2000000 6.998763
## 914          11              0.5000000            2.2000000 6.998763
## 921           0              0.3500000            0.0000000 4.730522
```

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.

```
df$cpg_z <- scale(df$cholesterol_per_gram)
subset(df,(df$cpg_z) >= 3)
```

```
##                                food.item weight_in_grams saturated_fat
## 46  BEEF LIVER; FRIED             3 OZ                85           2.5
## 59  CHICKEN LIVER; COOKED         1 LIVER             20           0.4
## 120 EGGS; RAW; YOLK               1 YOLK              17           1.6
## 168 EGGS; COOKED; FRIED           1 EGG               46           1.9
## 185 EGGS; RAW; WHOLE              1 EGG               50           1.6
## 186 EGGS; COOKED; POACHED         1 EGG               50           1.5
## 187 EGGS; COOKED; HARD-COOKED     1 EGG               50           1.6
## 190 EGGS; COOKED; SCRAMBLED/OMELET1 EGG               61           2.2
##     cholesterol saturated_fat_per_gram cholesterol_per_gram    sfpg.std
## 46          410             0.02941176             4.823529 -0.11728929
## 59          126             0.02000000             6.300000 -0.25961034
## 120         213             0.09411765            12.529412  0.86116796
## 168         211             0.04130435             4.586957  0.06254574
## 185         213             0.03200000             4.260000 -0.07815100
## 186         212             0.03000000             4.240000 -0.10839422
## 187         213             0.03200000             4.260000 -0.07815100
```

```
## 190         215             0.03606557             3.524590 -0.01667297
##         cpg_z
## 46   6.761927
## 59   8.947732
## 120 18.169910
## 168  6.411699
## 185  5.927664
## 186  5.898055
## 187  5.927664
## 190  4.838945
```