Lab 4 - In class exercise - Answers

Luis Navarro

1. Import the fastfood data set explored in class. Load the packages dplyr, ggplot2, and descriptr

```
# Clean the environment
rm(list=ls())

# Set your working directory
setwd("/Users/luisenriquenavarro/Library/CloudStorage/OneDrive-IndianaUniversity/V506/Fall24/Lab4")

# New Packages
if(!require(descriptr)){install.packages("descriptr")}
```

```
## Loading required package: descriptr
```

```
# Packages Required for the session
library(pacman)
p_load(dplyr, ggplot2, rmarkdown,
       descriptr,
       rio, here)

# Load fast food
fastfood <- rio::import(file = "fastfood.csv", header = TRUE) %>% tibble()
```

2. Following the code covered in class, write a function that takes as input a vector of data (say a variable from a data frame) The function should compute the mean, standard deviation, variance, and median of such variable. The output of the function should be a data frame with 4 columns: Mean, Median, SD, and Variance.

There are two potential answers for this. Using base R and summarize from dplyr.

```
# Base R
descriptive_stats <- function(data){

  # Compute the Descriptive Statistics using base R functions
  mean <- mean(data, na.rm = TRUE)
  median <- median(data, na.rm = TRUE)
  sd <- sd(data, na.rm = TRUE)
  var <- var(data, na.rm = TRUE)

  # Save output as data frame (tibble object)
  # note: I capitalize the variable names just to avoid confusion with the objects defined inside the f
  # you can assign them the same name as the objects and it works fine
  table <- tibble(
    Mean = mean,
    Median = median,
    SD = sd,
    VAR = var
    )

  # Return Table
  return(table)
}
```

```
# test the function
descriptive_stats(
  data = fastfood$calories
  )
```

```
## # A tibble: 1 x 4
##    Mean Median    SD    VAR
##   <dbl>  <int> <dbl>  <dbl>
## 1  531.    490  282. 79770.
```

```
# solution 2: using summarize from dplyr
desc_table2 <- function(data){
  # the key difference here is that data is not a vector, but a data frame
  # verify the data used as input is a data frame object (tibble).
  df <- tibble(variable = data)
  # create the table with desc stats
  table <- df %>%
    summarize(
      Mean = mean(variable, na.rm = TRUE),
      Median = median(variable, na.rm = TRUE),
      SD = sd(variable, na.rm = TRUE),
      VAR = var(variable, na.rm = TRUE)
    )

  return(table)
}
```

```
# test the function
desc_table2(
  data = fastfood$calories
)
```

```
## # A tibble: 1 x 4
##    Mean Median    SD    VAR
##   <dbl>  <int> <dbl>  <dbl>
## 1  531.    490  282. 79770.
```

3. Use your user written function to obtain the descriptive statistics of the variables: i) calories, ii) cholesterol, iii) protein, and iv) sugar. Use lapply and bind_rows to visualize the output in one data frame. Add a column with the variable names to identify the output of each row.

```r
# using base R
data_list <- list(fastfood$calories,
                  fastfood$cholesterol,
                  fastfood$protein,
                  fastfood$sugar)

# we can use lapply
data_summaries <- lapply(X = data_list,
                         FUN = desc_table2)

# merge them using bind_rows
data_summaries_table <- bind_rows(
  data_summaries[[1]] %>% mutate(var = "calories"),
  data_summaries[[2]] %>% mutate(var = "cholesterol"),
  data_summaries[[3]] %>% mutate(var = "protein"),
  data_summaries[[4]] %>% mutate(var = "sugar")
)

data_summaries
```

```
## [[1]]
## # A tibble: 1 x 4
##    Mean Median    SD    VAR
##   <dbl>  <int> <dbl>  <dbl>
## 1  531.    490  282. 79770.
##
## [[2]]
## # A tibble: 1 x 4
##    Mean Median    SD   VAR
##   <dbl>  <int> <dbl> <dbl>
## 1  72.5     60  63.2 3989.
##
## [[3]]
## # A tibble: 1 x 4
##    Mean Median    SD   VAR
##   <dbl>  <dbl> <dbl> <dbl>
## 1  27.9   24.5  17.7  313.
##
## [[4]]
## # A tibble: 1 x 4
##    Mean Median    SD   VAR
##   <dbl>  <int> <dbl> <dbl>
## 1  7.26      6  6.76  45.7
```

4. Use the function ds_summary_stats() from descriptr to compute the descriptive statistics of this new data frame. Compare the results with the output of your user written function.

```r
descriptr::ds_summary_stats(fastfood %>% select(calories, cholesterol, protein, sugar))
```

```
## ------------------------------ Variable: calories ------------------------------
##
##                              Univariate Analysis
```

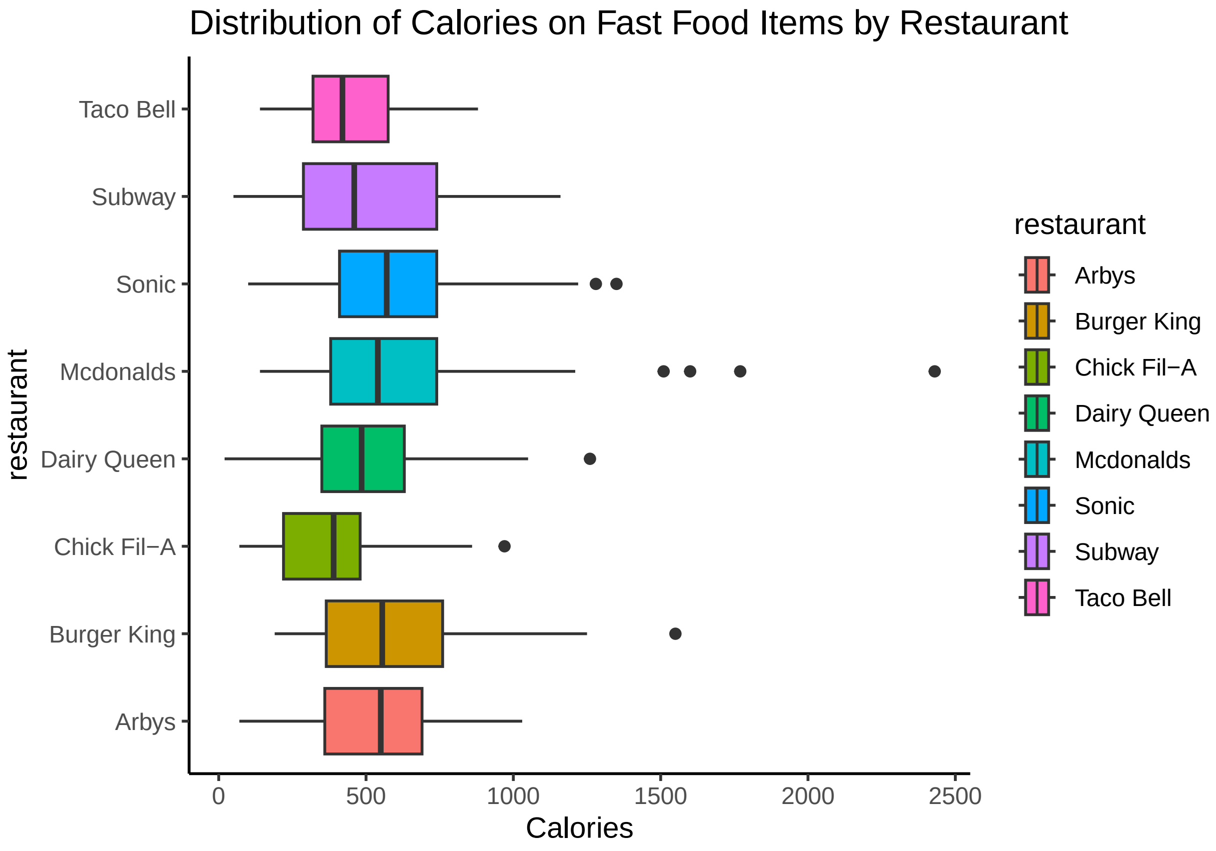
```
##
##  N                          515.00      Variance                 79770.18
##  Missing                      0.00      Std Deviation              282.44
##  Mean                       530.91      Range                     2410.00
##  Median                     490.00      Interquartile Range        360.00
##  Mode                       350.00      Uncorrected SS       186164000.00
##  Trimmed Mean               512.13      Corrected SS          41001871.07
##  Skewness                     1.41      Coeff Variation             53.20
##  Kurtosis                     4.72      Std Error Mean              12.45
##
##                                   Quantiles
##
##                 Quantile                                Value
##
##                Max                                     2430.00
##                99%                                     1340.20
##                95%                                     1033.00
##                90%                                      886.00
##                Q3                                       690.00
##                Median                                   490.00
##                Q1                                       330.00
##                10%                                      210.00
##                5%                                       170.00
##                1%                                       110.00
##                Min                                      20.00
##
##                                 Extreme Values
##
##                   Low                                    High
##
##   Obs                            Value       Obs                            Value
##   303                             20         40                             2430
##   393                             50         45                             1770
##   73                              70         48                             1600
##   188                             70         193                            1550
##   128                             100        39                             1510
##
##
##
## ---------------------------- Variable: cholesterol ----------------------------
##
##                           Univariate Analysis
##
##  N                        515.00      Variance               3989.24
##  Missing                    0.00      Std Deviation            63.16
##  Mean                      72.46      Range                   805.00
##  Median                    60.00      Interquartile Range      60.00
##  Mode                      40.00      Uncorrected SS       4754175.00
##  Trimmed Mean              65.69      Corrected SS         2050467.77
##  Skewness                   4.42      Coeff Variation          87.17
##  Kurtosis                  38.89      Std Error Mean            2.78
##
##                                 Quantiles
##
```

```
##                Quantile                              Value
##
##               Max                                    805.00
##               99%                                    282.20
##               95%                                    175.00
##               90%                                    130.00
##               Q3                                     95.00
##               Median                                 60.00
##               Q1                                     35.00
##               10%                                    20.00
##               5%                                     10.00
##               1%                                      0.00
##               Min                                     0.00
##
##                               Extreme Values
##
##                  Low                                  High
##
##   Obs                          Value       Obs                          Value
##   49                             0         193                           805
##   112                            0         40                            475
##   113                            0         206                           335
##   279                            0         39                            295
##   303                            0         45                            295
##
##
##
## ------------------------------ Variable: protein ------------------------------
##
##                          Univariate Analysis
##
##  N                       515.00      Variance                312.72
##  Missing                   1.00      Std Deviation            17.68
##  Mean                     27.89      Range                   185.00
##  Median                   24.50      Interquartile Range      20.00
##  Mode                     23.00      Uncorrected SS       560272.00
##  Trimmed Mean             26.27      Corrected SS         160425.90
##  Skewness                  2.81      Coeff Variation          63.40
##  Kurtosis                 16.49      Std Error Mean            0.78
##
##                                Quantiles
##
##               Quantile                              Value
##
##              Max                                   186.00
##              99%                                    96.61
##              95%                                    56.35
##              90%                                    46.70
##              Q3                                     36.00
##              Median                                 24.50
##              Q1                                     16.00
##              10%                                    12.00
##              5%                                     9.00
##              1%                                     6.00
##              Min                                    1.00
##
##                              Extreme Values
##
##                  Low                                High
##
##   Obs                         Value       Obs                         Value
##   302                           1         40                           186
##   392                           3         193                          134
##   188                           5         39                           115
##   233                           5         70                           103
##   135                           6         45                           98
##
##
##
## ------------------------------- Variable: sugar -------------------------------
##
##                         Univariate Analysis
##
##  N                      515.00      Variance                45.72
##  Missing                  0.00      Std Deviation            6.76
##  Mean                     7.26      Range                   87.00
##  Median                   6.00      Interquartile Range      6.00
##  Mode                     7.00      Uncorrected SS       50658.00
##  Trimmed Mean             6.63      Corrected SS         23497.61
##  Skewness                 4.61      Coeff Variation         93.10
##  Kurtosis                42.28      Std Error Mean           0.30
##
##                               Quantiles
##
##               Quantile                            Value
##
##              Max                                  87.00
##              99%                                  33.44
##              95%                                  16.00
##              90%                                  14.00
##              Q3                                    9.00
##              Median                                6.00
##              Q1                                    3.00
##              10%                                   1.00
##              5%                                    0.70
##              1%                                    0.00
##              Min                                   0.00
##
##                             Extreme Values
##
##                 Low                                High
##
##   Obs                        Value       Obs                        Value
##   35                           0         48                          87
##   41                           0         47                          52
##   42                           0         230                         37
##   43                           0         352                         36
##   44                           0         46                          35
```

5. Create a box plot for the distribution of calories on the fast food items across restaturants. On the x-axis show the distribution of "calories" and on the y-axis represent the restaurant. Use appropriate labels, coloring, and other graphics best practices.

```
boxplot_calories <- fastfood %>%
  ggplot(mapping = aes(x = calories, y = restaurant,
                       fill = restaurant)) +
  geom_boxplot() +
  labs(x = "Calories",
       title = "Distribution of Calories on Fast Food Items by Restaurant") +
  theme_classic()

boxplot_calories
```

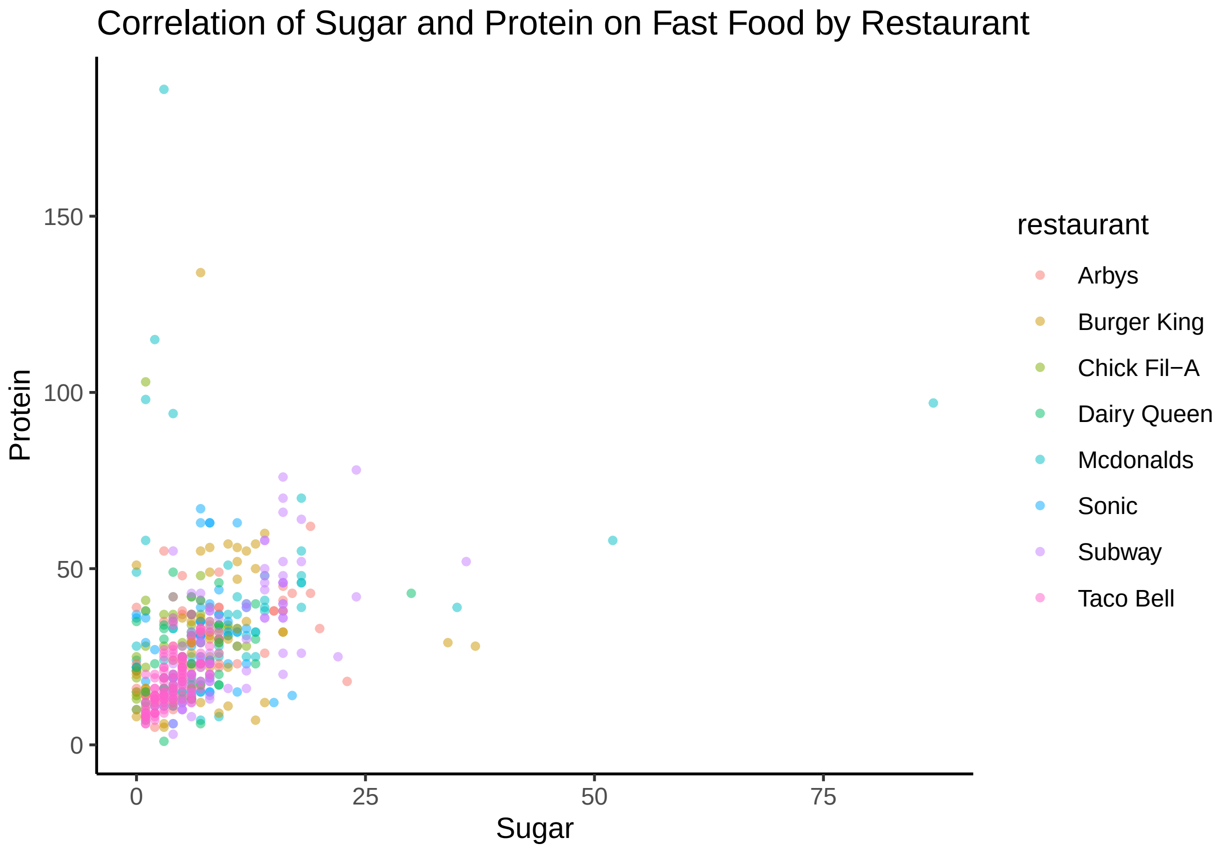


6. Create a scatter plot that shows the correlation between sugar (x-axis) and protein (y-axis) by restaurants on the same panel. Add a specific color for the observations of each restaurant. Use appropriate labels, coloring, and other graphics best practices. Save this graph in your environment. Hint: where in the aesthetic mapping should you specify that the graph will vary across restaurants?

```
scatter_sugar_protein <- fastfood %>%
  ggplot(mapping = aes(x = sugar, y = protein,
                       color = restaurant)) +
  geom_point(size = 1, alpha = 0.5) +
  labs(x = "Sugar", y = "Protein",
       title = "Correlation of Sugar and Protein on Fast Food by Restaurant")+
  theme_classic()

scatter_sugar_protein
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



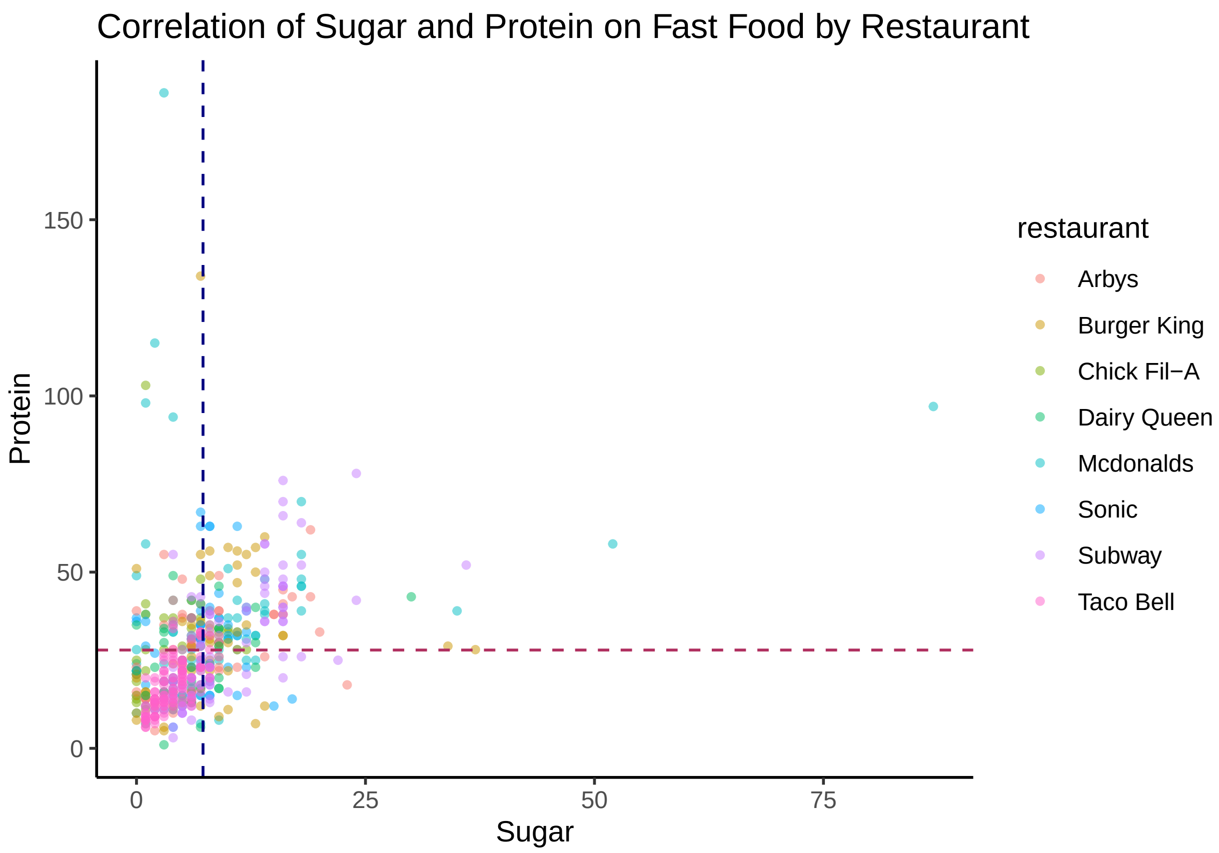
7. Add two lines to your previous scatterplot. A vertical line with the mean of variable sugar, and a horizontal line with the mean of variable cholesterol. Use geom_vline and geom_hline. For these functions you need to specify the xintercept and yintercept as parameters, respectively.

```
mean_sugar <- fastfood$sugar %>% mean(na.rm = TRUE)
mean_protein <- fastfood$protein %>% mean(na.rm = TRUE)

scatter_sugar_protein_improved <- scatter_sugar_protein +
  geom_vline(xintercept = mean_sugar, color = "navy", linetype = "dashed") +
  geom_hline(yintercept = mean_protein, color = "maroon", linetype = "dashed")

scatter_sugar_protein_improved
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



8. Export this graph using ggsave into your folder.

```
ggsave(filename = "scatter_protein_sugar.jpg",
       plot = scatter_sugar_protein_improved)
```

```
## Saving 6.5 x 4.5 in image
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```