Stat 420 - Homework 1

Fawad Khan

Fall 2025

Exercise 1 (Subsetting and Statistics)

(a)

```
library(ggplot2)
msleep
```

```
class(msleep)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

The msleep dataset is of object type tibble.

(b)

```
dim(msleep)
```

```
## [1] 83 11
```

There are 83 observations and 11 variables.

(c)

```
sum(is.na(msleep))
```

```
## [1] 136
```

```
sum(is.na(msleep$sleep_rem))
```

```
## [1] 22
```

```
mean(msleep$sleep_rem, na.rm = TRUE)
```

```
## [1] 1.87541
```

The mean hours of REM sleep is: 1.87541

(d)

```r
sd(msleep$brainwt, na.rm = TRUE )
```

```
## [1] 0.9764137
```

The standard deviation of brain weight of individuals is 0.9764137.

(e)

```r
msleep[which.max(msleep$sleep_rem), c('name', 'sleep_rem')]
```

```
## # A tibble: 1 x 2
##   name                 sleep_rem
##   <chr>                    <dbl>
## 1 Thick-tailed opposum       6.6
```

The observation with the most REM sleep is the Thick-tailed opposum.

(f)

```r
mean(msleep$bodywt[msleep$vore == "carni"], na.rm = TRUE)
```

```
## [1] 90.75111
```

The average bodyweight of carnivores in this dataset is 90.75111.

Exercise 2 (Plotting)

(a)

```r
library(MASS)
data(birthwt)
```

```r
?birthwt
colnames(birthwt)
```

The observations in this dataset represent factors associated with low infant birth weight. This includes factors such as the mothers age, smoking status during pregnancy, history of hypertension, and more.

(b)

```r
birthwt
class(birthwt)
```

This dataset is a dataframe object.

(c)

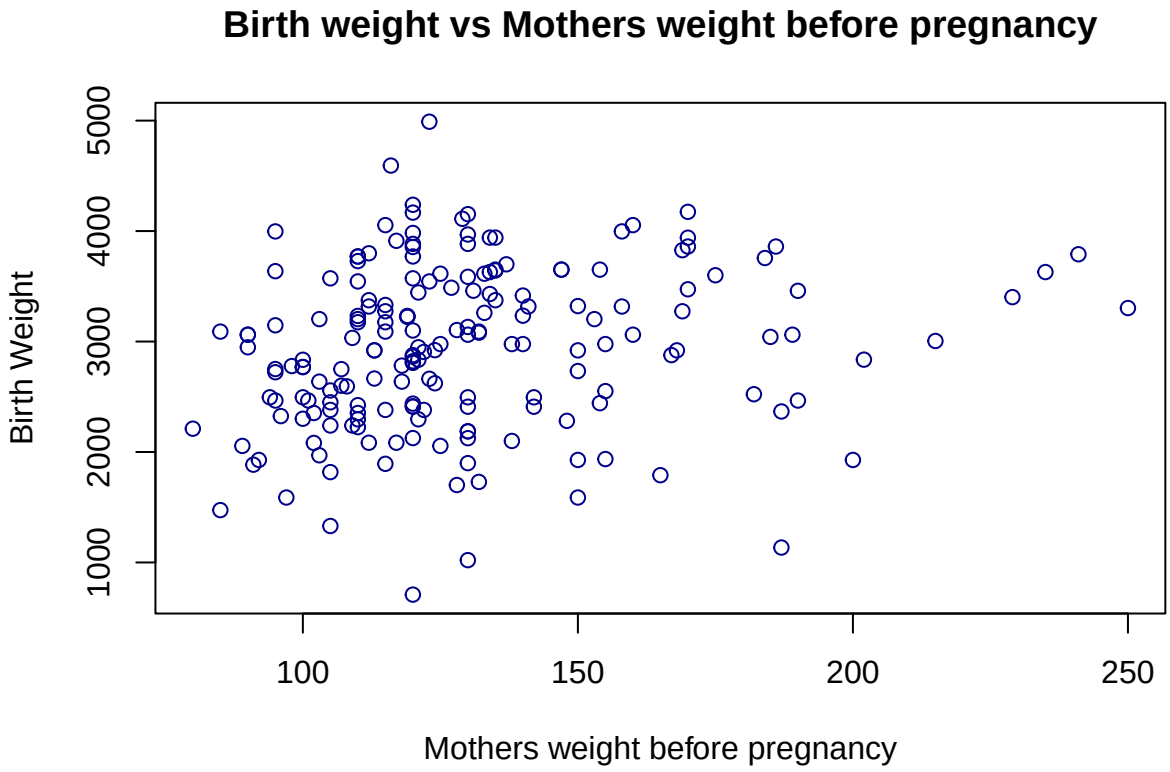
```r
dim(birthwt)
```

```
## [1] 189  10
```

This dataset contains 189 observations and 10 variables.

(d)

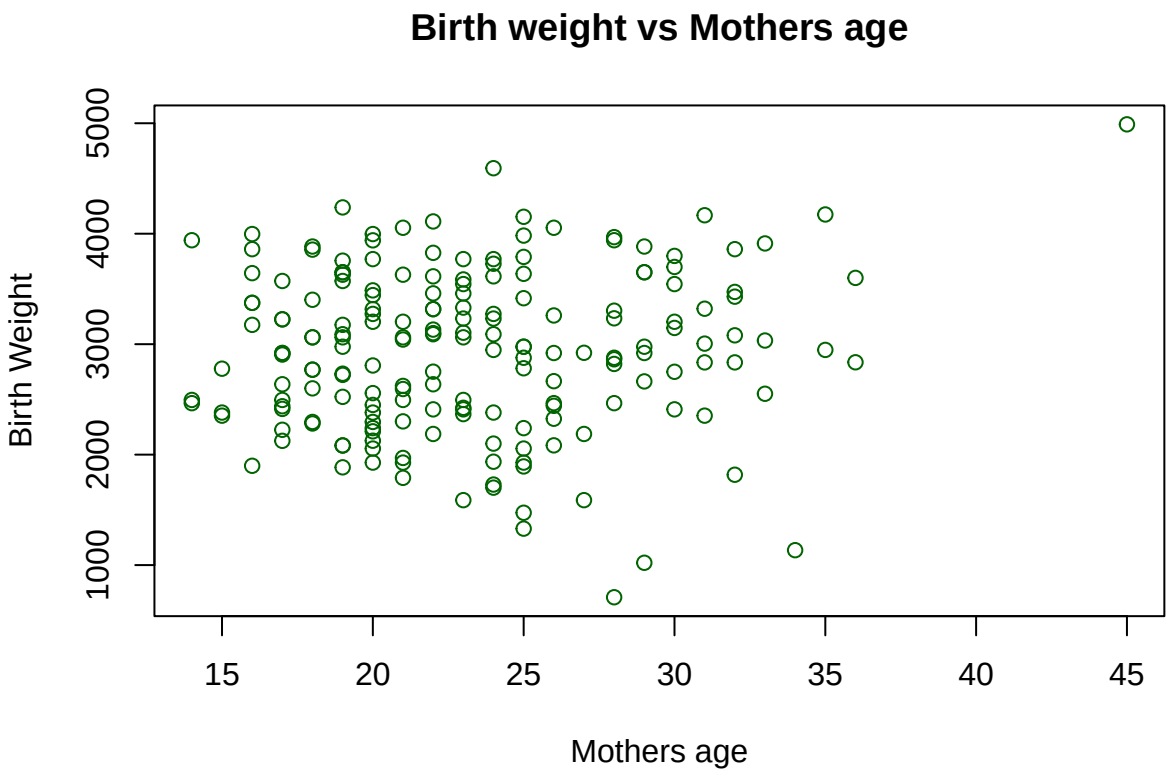
```r
plot(birthwt$lwt, birthwt$bwt,
  col = 'darkblue',
  xlab = 'Mothers weight before pregnancy',
  ylab = 'Birth Weight',
  main = 'Birth weight vs Mothers weight before pregnancy'

)
```



Based on the scatter plot, there is a slight upwards trend and so a positive relationship between the mothers weight before pregnancy and the birth weight. The higher the mothers weight before pregnancy, the higher the birth weight, although there is some variance.

(e)

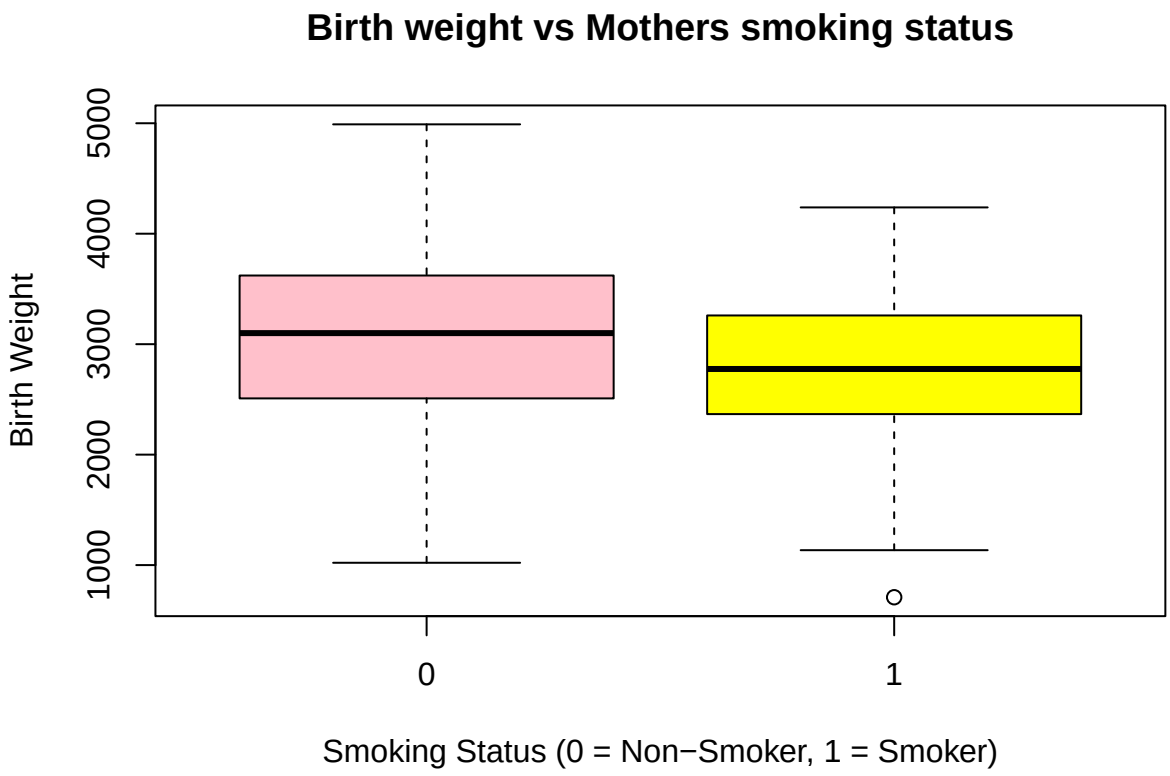
```r
plot(birthwt$age, birthwt$bwt,
     col = 'darkgreen',
     xlab = 'Mothers age',
     ylab = 'Birth Weight',
     main = 'Birth weight vs Mothers age'
)
```



Based on this plot, there does not seem to be a noticeable relationship between the two variables. The points appear scattered without a clear trend in any direction. This plot suggests that the mothers age is not a strong indicator of birth weight.

(f)

```
boxplot(birthwt$bwt ~ birthwt$smoke,
        col = c('pink', 'yellow'),
        xlab = 'Smoking Status (0 = Non-Smoker, 1 = Smoker)',
        ylab = 'Birth Weight',
        main = 'Birth weight vs Mothers smoking status'
)
```

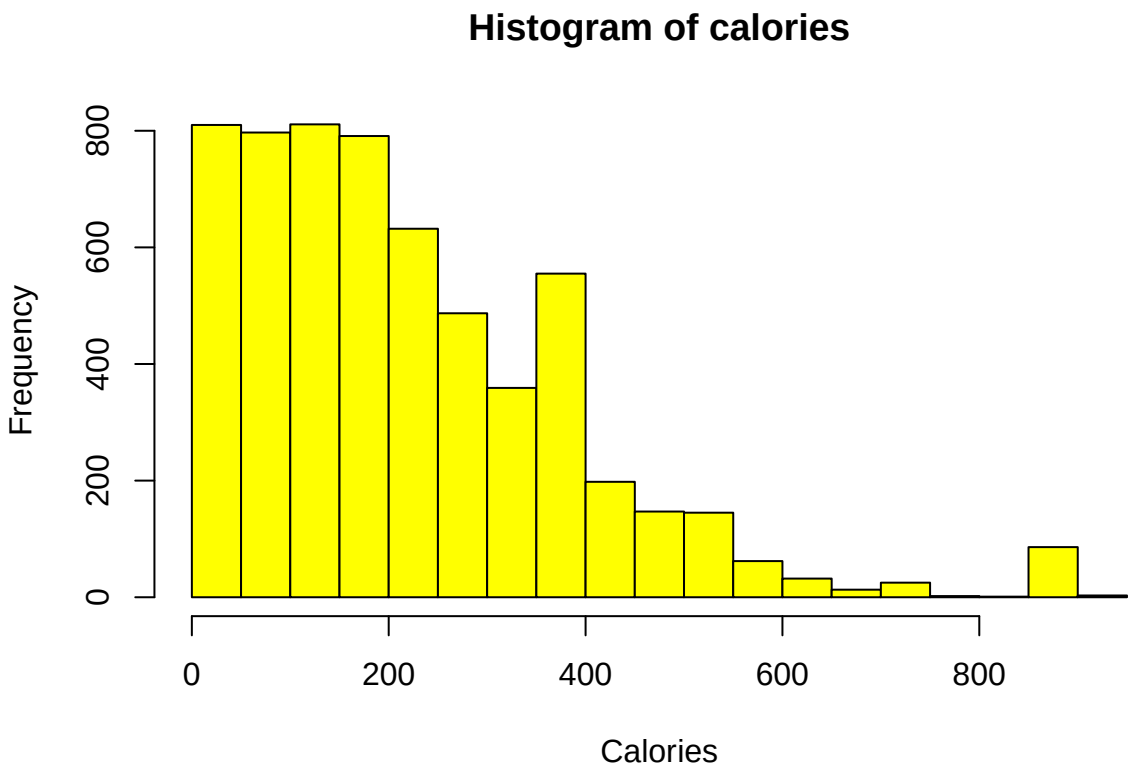


Based on this plot, we can see that the overall distribution of birth weight for smokers is shifted downwards, and the median birth weight of smokers is less than that of non-smokers. This suggests that smoking during pregnancy may lead to lower birth weight.

Exercise 3 (Importing data, more plotting)

(a)

```
hist(nutrition_2018$Calories,
     col = 'yellow',
     xlab = 'Calories',
     main = 'Histogram of calories'
)
```

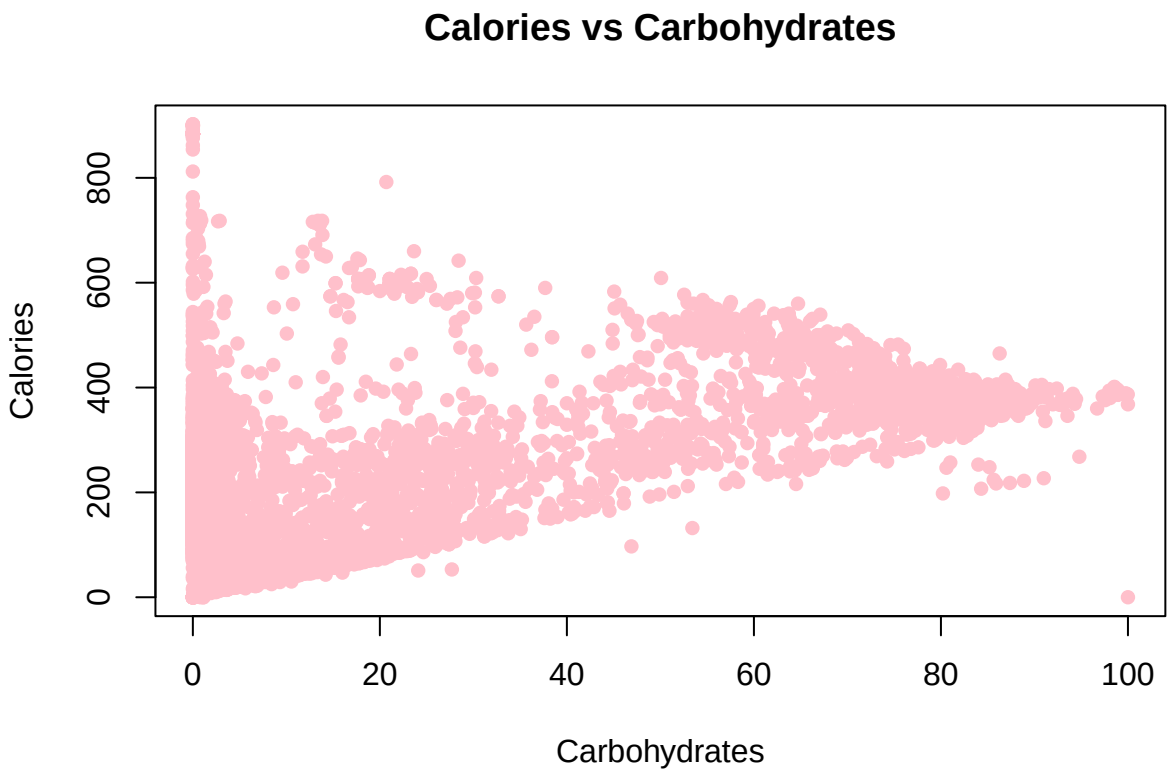


(b)

Based on this histogram, we can see that most foods are concentrated to the left and have a lower calorie count. There are definitely some outlines towards the right with a higher calorie count, and also some bins that have no representation. An explanation could be that the higher calorie foods are those that are high in sugar or fats, or are made with a lot of oil.

(c)

```
plot(nutrition_2018$Carbs, nutrition_2018$Calories,
     col = 'pink',
     xlab = 'Carbohydrates',
     ylab = 'Calories',
     main = 'Calories vs Carbohydrates',
     pch = 16
)
```

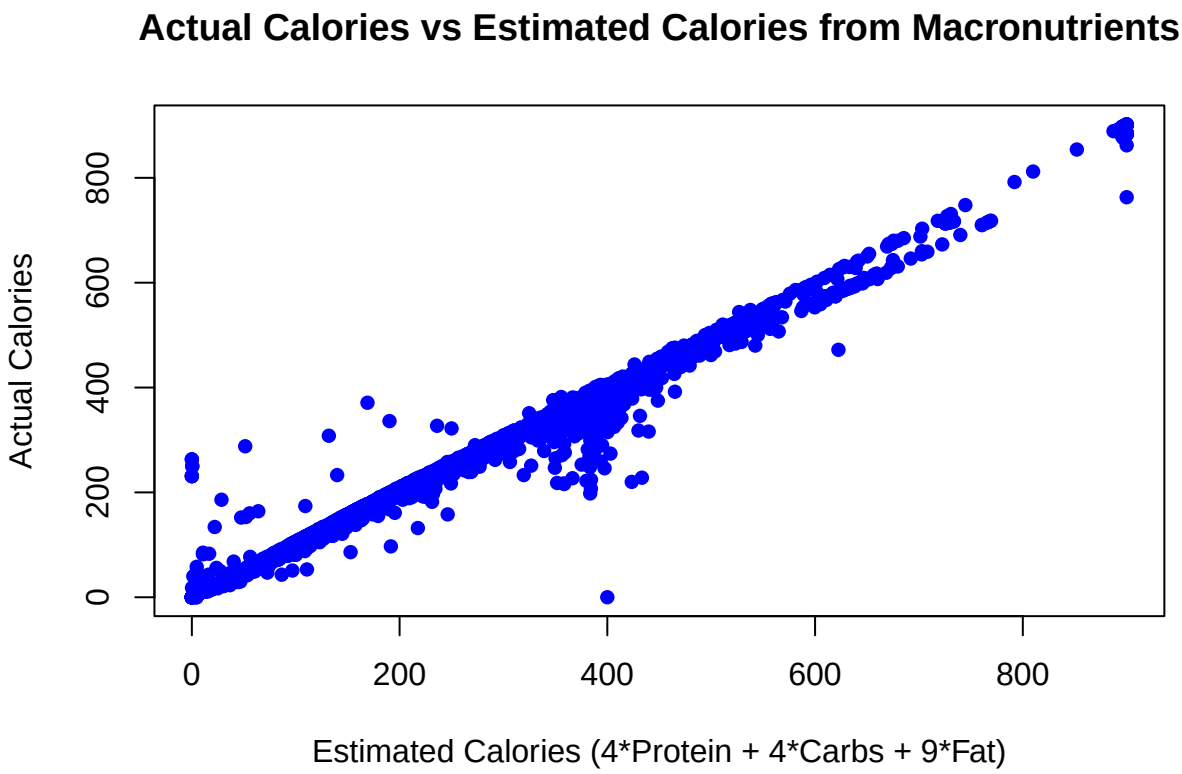


We can see a noticeable positive trend as carbohydrates increase, so do calories. However the same can be said of the variance, foods with lower carbohydrates seem be more spread out in terms of calories while those high in carbohydrates are more closely clustered together. It's interesting how much variance there is in foods with zero carbohydrates, they can be zero calories or go all the way up to 1000 calories. We can make a rough prediction of calories based on the number of carbohydrates but there are clearly more factors in play. To make a good, accurate prediction I don't think just carbohydrates is enough information.

(d)

```
nutrition_2018$estimatedCalories <- 4 * nutrition_2018$Protein +
  4 * nutrition_2018$Carbs + 9 * nutrition_2018$Fat
```

```
plot(nutrition_2018$estimatedCalories,nutrition_2018$Calories,
     col = 'blue',
     main = "Actual Calories vs Estimated Calories from Macronutrients",
     xlab = "Estimated Calories (4*Protein + 4*Carbs + 9*Fat)",
     ylab = "Actual Calories",
     pch = 16
)
```



The reason why this may not be a perfect line is because for one, we are only considering protein, carbs and fat, when there may be other factors in play such as fiber, sugar, starches, etc. Our dataset may also not be perfect, with the data being rounded or measured with some inaccuracies.

Exercise 4 (Writing and using functions)

(a)

```
sum_of_squares <- function(x) {
  return (sum(x^2))
}
```

```
a <- 1:20
sum_of_squares(a)
```

```
## [1] 2870
```

(b)

```
x <- c(11200, 7900, 7900, 4900, 4700)

x_bar <- mean(x)

deviations <- x - x_bar
```

```r
ssd <- sum_of_squares(deviations)

s2 <- ssd / (length(x) - 1)

s2
```

```
## [1] 7112000
```

```r
var(x)
```

```
## [1] 7112000
```

Both answers match.

(c)

```r
with(nutrition_2018, {
  residuals <- Calories - nutrition_2018$estimatedCalories
  sum_of_squares(residuals)
})
```

```
## [1] 2168424
```

Exercise 5 (More writing and using functions)

```r
set.seed(2025)
 random1 = rnorm(1000)
 random2 = runif(1000, min = 0, max = 1)
```

(a)

```r
list_extreme_values <- function(x, k = 2) {
  xbar <- mean(x, na.rm = TRUE)
  s <- sd(x, na.rm = TRUE)

  lower <- xbar - k * s
  upper <- xbar + k * s

  small <- x[x < lower]
  large <- x[x > upper]

  return(list(small = small, large = large))
}
```

```r
extreme_random1 <- list_extreme_values(random1, k = 2)
extreme_random1$small
```

```
##  [1] -2.450698 -2.317020 -2.849243 -2.234736 -2.624169 -2.461622 -2.329904
##  [8] -2.313391 -2.553343 -2.644750 -2.392542 -2.301849 -2.770171 -2.278110
## [15] -2.156053 -2.443789 -2.916665 -2.312980 -2.940893 -2.180501 -2.701685
```

```
extreme_random1$large
```

```
##  [1] 2.429019 2.852781 2.250047 2.104078 2.489617 2.835669 2.672551 2.100731
##  [9] 2.465668 2.395665 2.028216 2.021789 2.088209 2.019949 2.008089 3.270793
## [17] 3.152566 2.300967 2.478352 2.787236 2.094643 2.061602 2.047541 2.385399
## [25] 2.176122 2.081063
```

```
extreme_random2 <- list_extreme_values(random2, k = 2)
extreme_random2$small
```

```
## numeric(0)
```

```
extreme_random2$large
```

```
## numeric(0)
```

(b)

```
mean(list_extreme_values(random1, k = 2)$large)
```

```
## [1] 2.369168
```