Activity 4_2

Leonardo L Sanchez

2026-02-01

Question 1

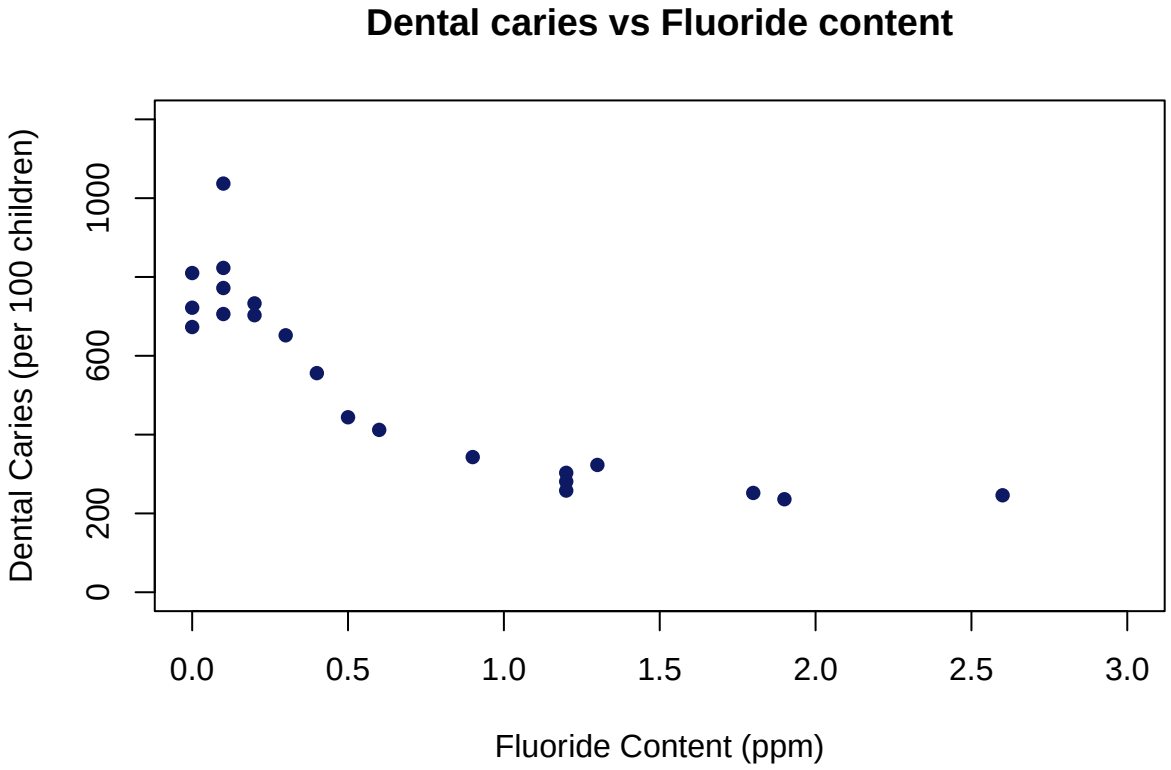
Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. The file water.Rdata contains the data from a study examining 7,257 children in 21 cities from the Flanders region in Belgium. The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved under the variable name fluoride; the number of dental caries per 100 children examined is saved under the name caries. The total dental caries number is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.

a) Construct a two-way scatterplot for these data, with fluoride as the x-variable and caries as the y-variable.

```
load("./water.Rdata")
View(water)
attach(water)
head(water)
```

```
##   fluoride caries
## 1      0.0    810
## 2      0.0    673
## 3      0.0    722
## 4      0.1    706
## 5      0.1    823
## 6      0.1   1037
```

```
plot(fluoride, caries, main = "Dental caries vs Fluoride content",
     xlab = "Fluoride Content (ppm)", ylab = "Dental Caries (per 100 children)",
     pch = 16, col = "#0D1A63",
     xlim=c(0,3), ylim=c(0,1200))
```



b) Do fluoride and caries appear to be positively or negatively associated? Explain your answer.

There is a relevant/significant negative correlation between fluoride content and dental caries. As fluoride levels increase, the number of dental caries decreases.

There is a rapid drop in dental caries and the number of dental caries remains low and relatively stable after fluoride content 1.0 . After reaching approximately 1.5 ppm, the graph looks plain and lateral behavior.

Question 2

This problem uses data from the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place between 2003 and 2006 in the Netherlands.
Body mass index (BMI) is a measure of body fat that is based on both height and weight. The World Health Organization and National Institutes for Health define a BMI of over 25.0 as overweight; this guideline is typically applied to adults in all age groups. However, a recent study has reported that individuals of ages 65 or older with the greatest mortality risk were those with BMI lower than 23.0, while those with BMI between 24.0 and 30.9 were at lower risk of mortality. These findings suggest that the ideal weight-for-height in older adults may not be the same as in younger adults. Explore the relationship between BMI (BMI) and age (age), using the same sample of 500 individuals from the prevend data.

a) Create a plot that shows the association between BMI and age. Based on the plot, comment briefly on the nature of the association.

https://umcgresearch.org/w/prevend

describe the natural course of chronic kidney disease; discover risk factors for developing chronic kidney disease; discover rick factors for developing cardiovascular disease, including heart failure and atrial fibrillation;

discover risk factors for developing type 2 diabetes; investigate new methods for prevention; advocate for the impact of chronic kidney disease.

```
load("./prevend.rda")
View(prevend)
attach(prevend)
head(prevend)
```

```
##   Casenr Age Gender Ethnicity Education RFFT VAT CVD DM Smoking Hypertension
## 1      1  35      0         0         3   58  11   0  0       1            0
## 2      2  35      0         0         3   82  11   0  0       0            0
## 3      3  35      0         0         2  105  10   0  0       1            0
## 4      4  35      0         0         2   39  12   0  0       0            0
## 5      5  35      0         0         3   94  -1   0  0       0            0
## 6      6  35      0         0         1   40   9   0  0       0            0
##        BMI   SBP  DBP  MAP      eGFR Albuminuria.1 Albuminuria.2 Chol  HDL
## 1 24.61521 116.0 64.5 78.5  68.22967             0             1 5.50 0.94
## 2 21.23410 117.5 61.0 81.0 104.55268             0             0 3.65 0.87
## 3 29.24149 132.5 79.0 98.5  98.54159             1             2 6.93 1.14
## 4 29.15877 130.5 79.5 98.5 113.09780             0             0 3.95 0.98
## 5 29.58222 118.5 71.5 88.0  90.64893             0             1 4.58 0.92
## 6 27.42382 124.5 69.0 89.0  81.80339             0             0 5.64 1.10
##   Statin Solubility Days Years DDD FRS         PS PSquint GRS Match_1 Match_2
## 1      0          2   -1    -1   0   7 0.08101016       2   1      -1      -1
## 2      0          2   -1    -1   0   2 0.06069326       1   0      -1      -1
## 3      0          2   -1    -1   0  11 0.17047626       3   1      -1      -1
## 4      0          2   -1    -1   0   4 0.09005986       2   0      -1      -1
## 5      0          2   -1    -1   0   2 0.07644714       2   0      -1      -1
## 6      0          2   -1    -1   0   3 0.10591498       2   0      -1      -1
```

```
model <- lm(BMI ~ Age, data = prevend)

summary(model)
```

```
##
## Call:
## lm(formula = BMI ~ Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.457  -2.968  -0.639   2.269  34.196
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.807168   0.325045   70.17   <2e-16 ***
## Age          0.073040   0.005818   12.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.316 on 4093 degrees of freedom
## Multiple R-squared:  0.03708,    Adjusted R-squared:  0.03684
## F-statistic: 157.6 on 1 and 4093 DF,  p-value: < 2.2e-16
```

c) Write the equation of the linear model.

```
cef= coef(model)
cef
```

```
## (Intercept)         Age
## 22.80716782  0.07304024
```

```
b0 <- coef(model)[1]
b1 <- coef(model)[2]

cat(
  "E(BMI) =",
  round(b0, 5),
  "+",
  round(b1, 5),
  "*Age"
)
```

```
## E(BMI) = 22.80717 + 0.07304 *Age
```

d) Interpret the slope and intercept values in the context of the data. Comment on whether the intercept value has any interpretive meaning in this setting.

$\beta_0 = 22.80716782$: When age is 0, the predicted BMI is 22.81. However, this value is not meaningful in practice.
$\beta_1 = 0.07304024$: For each additional year of age, the BMI increases by 0.073 on average.

e) Is it valid to use the linear model to estimate BMI for an individual who is 30 years old? Explain your answer.

```
range(prevend$Age)
```

```
## [1] 35 82
```

the data ranges from 35 to 82 years, predicting BMI for a 30-year-old involves extrapolation outside low range from the experimental region. So, estimating BMI for a 30 year old using this model is not recommended.

f) According to the linear model, estimate the average BMI for an individual who is 60 years old.

```
predict(model, data.frame(Age = 60))
```

```
##        1
## 27.18958
```

The prediction gives an estimated BMI of approximately 27.19.

g) Based on the linear model, how much does BMI differ, on average, between an individual who is 70 years old versus an individual who is 50 years old?

```
predictdiff=predict(model, data.frame(Age = 70))-predict(model, data.frame(Age = 50))

print(predictdiff)
```

```
##        1
## 1.460805
```

```
print(coef(model)[2] * (70 - 50))
```

```
##      Age
## 1.460805
```

differ E(y) = 1.460805

h) Conduct a formal hypothesis test of no association between BMI and age, at the alpha= 0.5 significance level. Summarize your conclusions.

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

```
summary(model)$coefficients["Age", "Pr(>|t|)"]
```

```
## [1] 1.698688e-35
```

Since the p-value 1.698688e-35 is far less than 0.05, we reject the null hypothesis. This indicates that age is significantly associated with BMI.

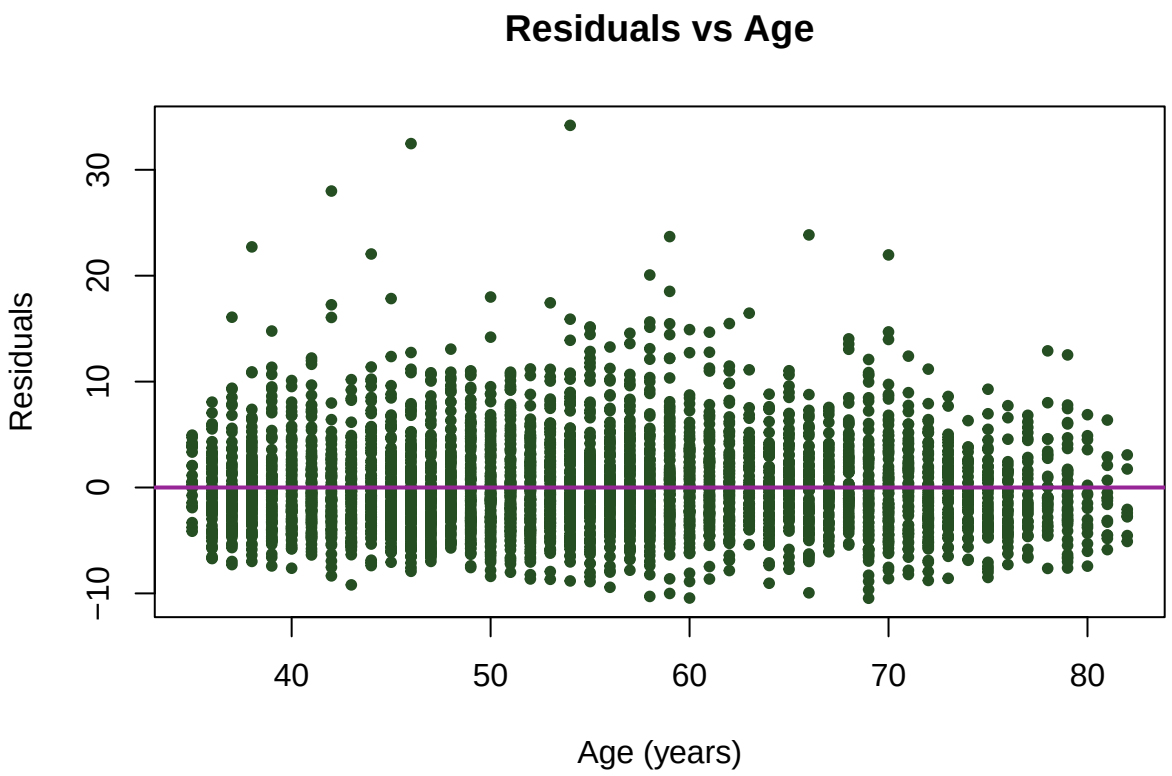
i) Report the R2 of the linear model relating BMI and age. Based on the R2 value, briefly comment on whether you think the estimated average BMI values calculated in part b) are accurate.

R-squared: 0.03684

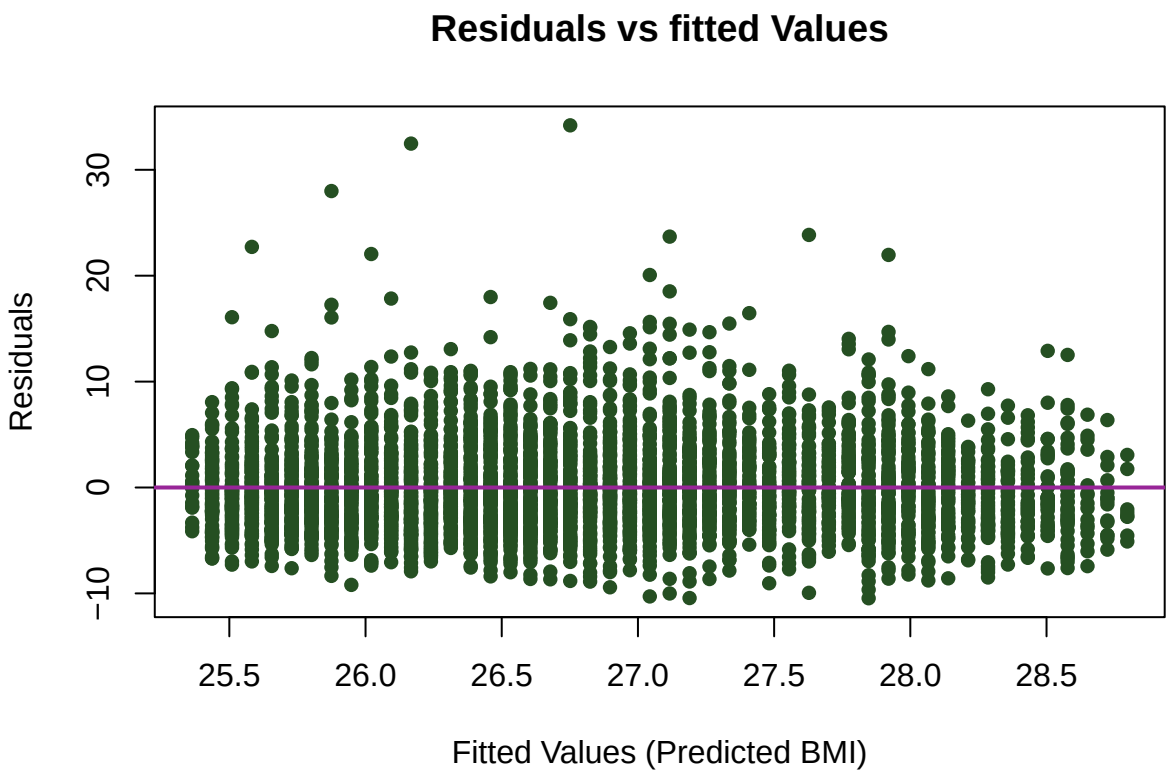
R-squared indicates that only 3.7% of the variability in BMI is explained by age. This is a very low proportion, meaning that age alone is not a strong predictor of BMI. Although the model is statistically significant (p < 2.2e-16), the low R-squared suggests it has limited practical usefulness in predicting BMI based solely on age.

j) Create residual plots to assess the model assumptions of linearity, constant variability, and normally distributed residuals. In your assessment of whether an assumption is reasonable, be sure to clearly reference and interpret relevant features of the appropriate plot.

```
# Residuals vs Age
plot(prevend$Age, resid(model),
     main = "Residuals vs Age",
     xlab = "Age (years)", ylab = "Residuals",
     pch = 20, col = "#254F22")
abline(h = 0, col = "#982598", lwd = 2) # Add reference line at 0
```




```r
# Residuals vs Fitted Values
plot(fitted(model), resid(model),
     main = "Residuals vs fitted Values",
     xlab = "Fitted Values (Predicted BMI)", ylab = "Residuals",
     pch = 16, col = "#254F22")
abline(h = 0, col = "#982598", lwd = 2) # Add reference line at 0
```

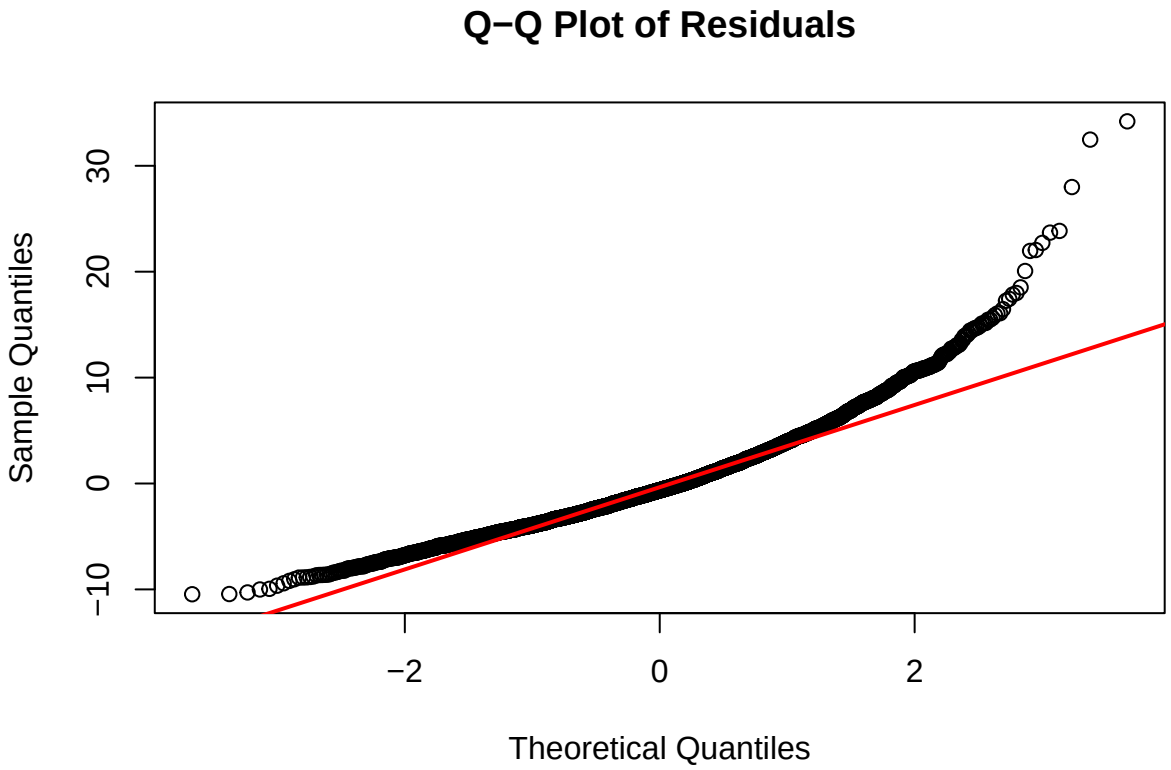


Interpretation

Residuals vs Age Plot: The residuals are randomly scattered around without a clear trend zero, the zero line means means the model predicted the value exactly.

the graph depicts that the relationship between BMI and age is approximately linear. There are some outliers, but no clear violation of linearity is observed. Residuals vs Fitted Values Plot: The points are evenly spread around zero.
The residual plots do not indicate strong violations of linearity or constant variance assumptions. The model reasonably meets the assumptions for a simple linear regression, but other predictors might improve model accuracy.

```
qqnorm(resid(model), main = "Q-Q Plot of Residuals")
qqline(resid(model), col = "red", lwd = 2)
```



The residuals do not perfectly follow a normal distribution, as there is a noticeable departure at the tails. This suggests potential outliers or non-normality, which could affect inference and prediction accuracy.

Question 3

This problem uses data from the National Health and Nutrition Examination Survey (NHANES), a survey conducted annually by the US Centers for Disease Control (CDC). The data can be treated as if it were a simple random sample from the American population. The dataset nhanes.samp.adult.500 contains data for 500 participants ages 21 years or older that were randomly sampled from the complete NHANES dataset that contains 10,000 observations. Regular physical activity is important for maintaining a healthy weight, boosting mood, and reducing risk for diabetes, heart attack, and stroke. In this problem, you will be exploring the relationship between weight (Weight) and physical activity (PhysActive) using the data in nhanes.samp.adult.500. Weight is measured in kilograms. The variable PhysActive is coded Yes if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities, and No if otherwise.

```
load("./nhanes.samp.adult.500.rda")
View(nhanes.samp.adult.500)
attach(nhanes.samp.adult.500)
```

```
## The following objects are masked from prevend:
##
##     Age, BMI, Education, Gender
```

```
head(nhanes.samp.adult.500)
```

```
##         ID SurveyYr Gender Age AgeDecade AgeMonths Race1 Race3      Education
## 5514 63106  2011_12   male  50     50-59        NA White White 9 - 11th Grade
## 7882 67820  2011_12 female  47     40-49        NA Black Black   College Grad
## 2619 57178  2009_10   male  46     40-49       561 White  <NA>   Some College
## 8361 68693  2011_12   male  28     20-29        NA White White   Some College
```

```
## 8725 69465  2011_12 female  50     50-59        NA White White   College Grad
## 4692 61505  2009_10   male  39     30-39       471 Black  <NA>   Some College
##      MaritalStatus    HHIncome HHIncomeMid Poverty HomeRooms HomeOwn       Work
## 5514      Divorced 10000-14999       12500    0.95         7     Own NotWorking
## 7882     Separated 35000-44999       40000    1.74         6    Rent    Working
## 2619       Married  more 99999      100000    4.99         6     Own    Working
## 8361  NeverMarried  more 99999      100000    4.14         8     Own    Working
## 8725      Divorced 35000-44999       40000    2.16        10     Own NotWorking
## 4692  NeverMarried        <NA>          NA      NA         4    Rent    Working
##      Weight Length HeadCirc Height   BMI BMICatUnder20yrs      BMI_WHO Pulse
## 5514   82.8     NA       NA  172.2 27.90             <NA> 25.0_to_29.9    58
## 7882   79.9     NA       NA  164.8 29.40             <NA> 25.0_to_29.9    70
## 2619   73.7     NA       NA  170.5 25.35             <NA> 25.0_to_29.9    74
## 8361   80.9     NA       NA  177.1 25.80             <NA> 25.0_to_29.9    58
## 8725   70.5     NA       NA  161.8 26.90             <NA> 25.0_to_29.9    76
## 4692  120.0     NA       NA  190.9 32.93             <NA>    30.0_plus    58
##      BPSysAve BPDiaAve BPSys1 BPDia1 BPSys2 BPDia2 BPSys3 BPDia3 Testosterone
## 5514      125       86    122     88    124     86    126     86       525.37
## 7882      121       68    124     66    120     66    122     70         5.98
## 2619      120       74    120     70    118     74    122     74           NA
## 8361      132       74    134     72    130     72    134     76       653.19
## 8725      152      103    144    104    150    106    154    100         8.17
## 4692      148       88    150     94    148     92    148     84           NA
##      DirectChol TotChol UrineVol1 UrineFlow1 UrineVol2 UrineFlow2 Diabetes
## 5514       1.29    5.07       244      1.683        NA         NA       No
## 7882       1.22    3.70        65      0.442        NA         NA       No
## 2619       1.40    6.03       105      0.682        NA         NA       No
## 8361       1.84    4.55        51      0.464        NA         NA       No
## 8725       2.43    5.92        30      1.304       114      1.118       No
## 4692         NA      NA        NA         NA        NA         NA       No
##      DiabetesAge HealthGen DaysPhysHlthBad DaysMentHlthBad LittleInterest
## 5514          NA      Fair               5              30          Most
## 7882          NA     Vgood               2               5          None
## 2619          NA     Vgood               0               0          None
## 8361          NA Excellent               0               0          None
## 8725          NA     Vgood               0               0          None
## 4692          NA      <NA>              NA              NA          <NA>
##      Depressed nPregnancies nBabies Age1stBaby SleepHrsNight SleepTrouble
## 5514   Several           NA      NA         NA             4          Yes
## 7882   Several            2       2         21             6           No
## 2619      None           NA      NA         NA             5           No
## 8361      None           NA      NA         NA             7           No
## 8725      None            3       3         27             6          Yes
## 4692      <NA>           NA      NA         NA             6           No
##      PhysActive PhysActiveDays  TVHrsDay CompHrsDay TVHrsDayChild
## 5514         No              3      2_hr      0_hrs           NA
## 7882        Yes             NA      3_hr       2_hr           NA
## 2619        Yes              3      <NA>       <NA>           NA
## 8361        Yes             NA 0_to_1_hr       4_hr           NA
## 8725        Yes             NA      1_hr       1_hr           NA
## 4692        Yes              4      <NA>       <NA>           NA
##      CompHrsDayChild Alcohol12PlusYr AlcoholDay AlcoholYear SmokeNow Smoke100
## 5514              NA             Yes          1          24       No      Yes
## 7882              NA             Yes          2           3     <NA>       No
```

```
## 2619              NA             Yes          NA            0     <NA>       No
## 8361              NA              No          NA            0     <NA>       No
## 8725              NA             Yes           1          364     <NA>       No
## 4692              NA            <NA>          NA           NA     <NA>       No
##       Smoke100n SmokeAge Marijuana AgeFirstMarij RegularMarij AgeRegMarij
## 5514     Smoker       18        No            NA           No          NA
## 7882 Non-Smoker       NA       Yes            19          Yes          20
## 2619 Non-Smoker       NA       Yes            14          Yes          16
## 8361 Non-Smoker       NA        No            NA           No          NA
## 8725 Non-Smoker       NA        No            NA           No          NA
## 4692 Non-Smoker       NA      <NA>            NA         <NA>          NA
##      HardDrugs SexEver SexAge SexNumPartnLife SexNumPartYear SameSex
## 5514        No     Yes     16              26              2      No
## 7882        No     Yes     17              10              2      No
## 2619       Yes     Yes     14              50              1      No
## 8361        No      No     NA               0              0      No
## 8725        No     Yes     17               4              1      No
## 4692      <NA>    <NA>     NA              NA             NA    <NA>
##      SexOrientation PregnantNow
## 5514   Heterosexual        <NA>
## 7882   Heterosexual        <NA>
## 2619   Heterosexual        <NA>
## 8361   Heterosexual        <NA>
## 8725   Heterosexual        <NA>
## 4692           <NA>        <NA>
```

a) Explore the data.

i.Identify how many individuals are physically active.

```
table(nhanes.samp.adult.500$PhysActive)
```

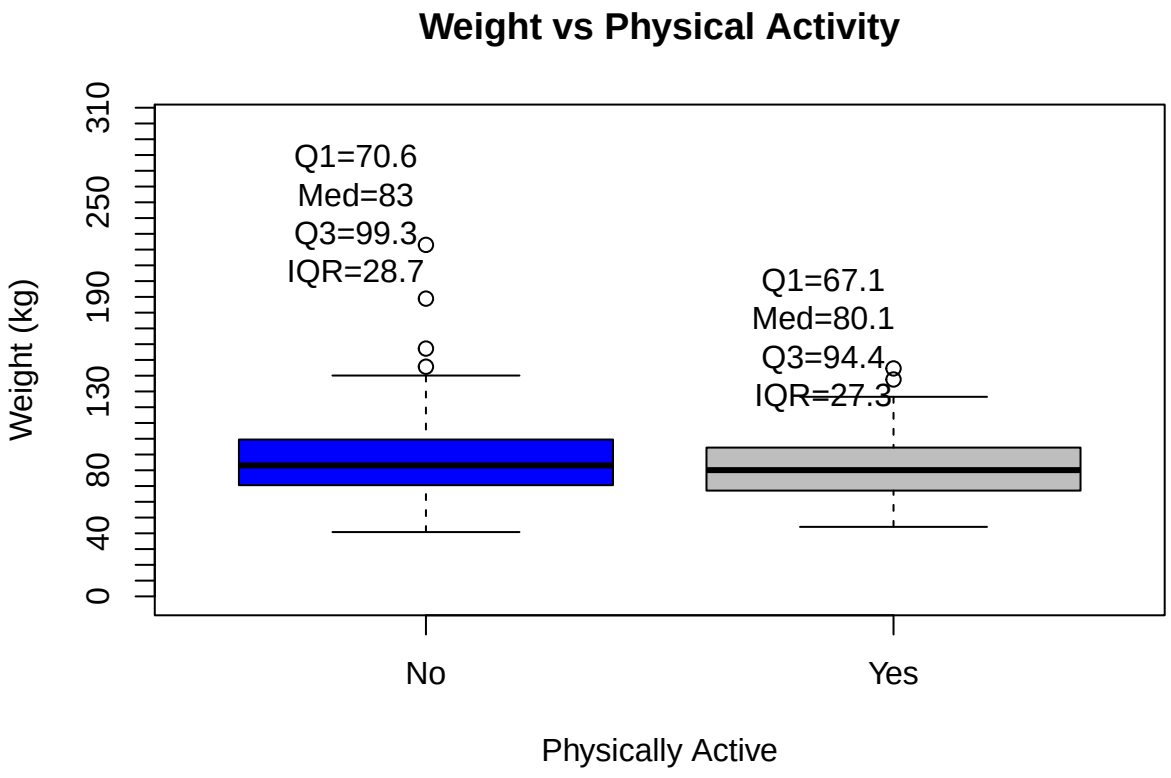
```
##
##  No Yes
## 250 250
```

250 individuals are physically active

ii. Create a plot that shows the association between weight and physical activity. Describe what you see.

```
boxplot(
  Weight ~ PhysActive,
  data = nhanes.samp.adult.500,
  main = "Weight vs Physical Activity",
  xlab = "Physically Active",
  ylab = "Weight (kg)",
  col = c("blue", "grey"),
  ylim = c(0, 300),
  yaxt = "n"   # turn off default y-axis
)

# Add y-axis with ticks every 25
axis(2, at = seq(0, 350, by = 10))
```

```r
groups <- levels(nhanes.samp.adult.500$PhysActive)

for (i in seq_along(groups)) {

  x <- nhanes.samp.adult.500$Weight[
    nhanes.samp.adult.500$PhysActive == groups[i]
  ]

  q1  <- quantile(x, 0.25, na.rm = TRUE)
  q2  <- median(x, na.rm = TRUE)
  q3  <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- IQR(x, na.rm = TRUE)

  # place text ABOVE all data points
  y_pos <- max(x, na.rm = TRUE) +20

  text(
     i - 0.15, y_pos,
    labels = paste0(
      "Q1=", round(q1,1), "\n",
      "Med=", round(q2,1), "\n",
      "Q3=", round(q3,1), "\n",
      "IQR=", round(iqr,1)
    ),
    cex = 1
  )
}
```



Interpretation The median weight appears slightly lower for physically active individuals ("Yes") compared to those who are not physically active ("No"). This suggests that physically active individuals tend to have slightly lower body weight on average.
The interquartile range is similar for both groups, meaning the weight distribution is comparable. Both groups show a wide range of weights.
There are several outliers in both groups, representing individuals with exceptionally high body weight (>150 kg).