Module-9-final-project.R

HI

2024-09-16

#installing required libraries  
library(ggplot2)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

#loading the data  
url <- "https://raw.githubusercontent.com/HackBio-Internship/public\_datasets/main/R/nhanes.csv"  
nhanes\_data <- read.csv(url, header = TRUE, sep = ",")  
  
#show first few rows  
head(nhanes\_data)

## id Gender Age Race Education MaritalStatus RelationshipStatus Insured  
## 1 62163 male 14 Asian <NA> <NA> <NA> Yes  
## 2 62172 female 43 Black High School NeverMarried Single Yes  
## 3 62174 male 80 White College Grad Married Committed Yes  
## 4 62174 male 80 White College Grad Married Committed Yes  
## 5 62175 male 5 White <NA> <NA> <NA> Yes  
## 6 62176 female 34 White College Grad Married Committed Yes  
## Income Poverty HomeRooms HomeOwn Work Weight Height BMI Pulse BPSys  
## 1 100000 4.07 6 Rent <NA> 49.4 168.9 17.3 72 107  
## 2 22500 2.02 4 Rent NotWorking 98.6 172.0 33.3 80 103  
## 3 70000 4.30 7 Own NotWorking 95.8 168.1 33.9 56 97  
## 4 70000 4.30 7 Own NotWorking 95.8 168.1 33.9 56 97  
## 5 12500 0.39 7 Rent <NA> 23.9 119.8 16.7 NA NA  
## 6 100000 5.00 8 Own NotWorking 68.7 171.6 23.3 92 107  
## BPDia Testosterone HDLChol TotChol Diabetes DiabetesAge nPregnancies nBabies  
## 1 37 274.95 1.14 3.98 No NA NA NA  
## 2 72 47.53 1.89 4.37 No NA 3 2  
## 3 39 642.82 1.40 5.25 No NA NA NA  
## 4 39 642.82 1.40 5.25 No NA NA NA  
## 5 NA NA NA NA No NA NA NA  
## 6 69 21.11 1.42 4.42 No NA 5 2  
## SleepHrsNight PhysActive PhysActiveDays AlcoholDay AlcoholYear SmokingStatus  
## 1 NA No 1 NA NA <NA>  
## 2 8 No 2 3 104 Current  
## 3 9 No 7 NA 0 Never  
## 4 9 No 5 NA 0 Never  
## 5 NA <NA> 7 NA NA <NA>  
## 6 7 Yes 5 2 104 Never

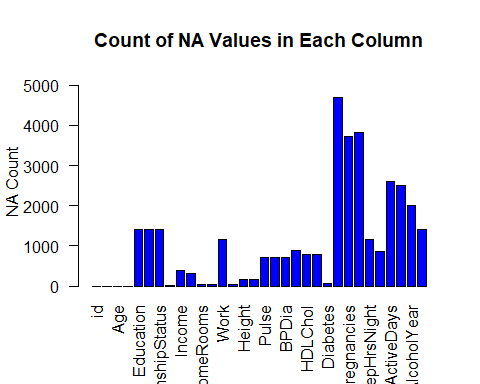
names(nhanes\_data)

## [1] "id" "Gender" "Age"   
## [4] "Race" "Education" "MaritalStatus"   
## [7] "RelationshipStatus" "Insured" "Income"   
## [10] "Poverty" "HomeRooms" "HomeOwn"   
## [13] "Work" "Weight" "Height"   
## [16] "BMI" "Pulse" "BPSys"   
## [19] "BPDia" "Testosterone" "HDLChol"   
## [22] "TotChol" "Diabetes" "DiabetesAge"   
## [25] "nPregnancies" "nBabies" "SleepHrsNight"   
## [28] "PhysActive" "PhysActiveDays" "AlcoholDay"   
## [31] "AlcoholYear" "SmokingStatus"

#Checking the NA values in the dataset  
sum(is.na(nhanes\_data))

## [1] 33991

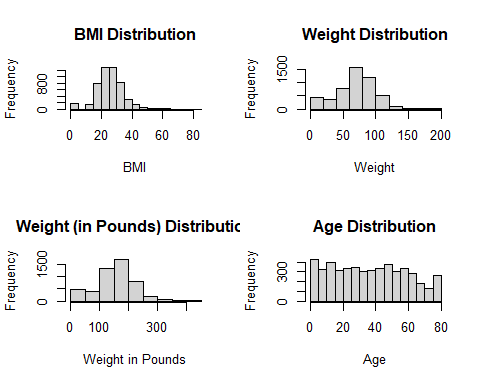
#checking the NA values in each columns  
na\_count <- colSums(is.na(nhanes\_data))  
  
#Visualising the count of NA values per column  
barplot(na\_count,   
 main = "Count of NA Values in Each Column",  
 #xlab = "Columns",  
 ylab = "NA Count",  
 ylim = c(0, max(na\_count) \* 1.1),  
 names.arg = names(na\_count),  
 las = 2,  
 col = "blue")



# deleting all NA  
#nhanes\_cleaned <- na.omit(nhanes\_data)  
  
#Replacing NA with 0  
nhanes\_replaced <- nhanes\_data  
nhanes\_replaced[is.na(nhanes\_replaced)] <- 0  
  
# verifying the NA counts After being replaced  
sum(is.na(nhanes\_replaced))

## [1] 0

#Adding Weight in pounds (Weight \* 2.2) as a new variable  
nhanes\_replaced$Weight\_pounds <- nhanes\_replaced$Weight \* 2.2  
  
# Plotting histograms for4 different variables in a 2 \* 2 grid  
par(mfrow=c(2,2))  
  
hist(nhanes\_replaced$BMI, main="BMI Distribution", xlab="BMI")  
hist(nhanes\_replaced$Weight, main="Weight Distribution", xlab="Weight")  
hist(nhanes\_replaced$Weight\_pounds, main="Weight (in Pounds) Distribution",   
 xlab="Weight in Pounds")  
hist(nhanes\_replaced$Age, main="Age Distribution", xlab="Age")



# summarising the data to get the mean 60-second pulse rate for all participants  
summary(nhanes\_replaced$Pulse)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 60.00 70.00 63.06 80.00 136.00

#mean(nhanes\_replaced$Pulse == 60)  
  
  
# the range of values for diastolic blood pressure in all participants  
min(nhanes\_replaced$BPDia)

## [1] 0

max(nhanes\_replaced$BPDia)

## [1] 116

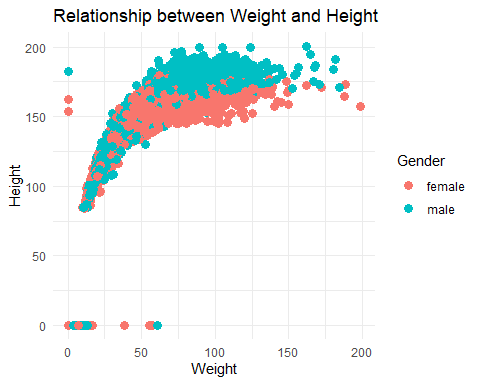
# checking the variance and standard deviation of the income column  
var(nhanes\_replaced$Income)

## [1] 1264147754

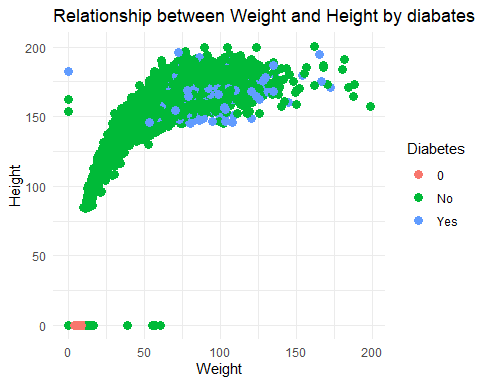
sd(nhanes\_replaced$Income)

## [1] 35554.86

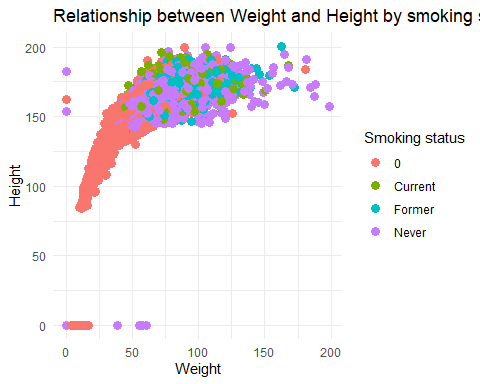
# Scatter plot with points colored by gender  
ggplot(nhanes\_replaced, aes(x = Weight, y = Height, color = Gender)) +  
 geom\_point(size = 3) +  
 labs(title = "Relationship between Weight and Height",  
 x = "Weight", y = "Height",  
 color = "Gender") +  
 theme\_minimal()



# Scatter plot with points colored by diabetes  
ggplot(nhanes\_replaced, aes(x = Weight, y = Height, color = Diabetes)) +  
 geom\_point(size = 3) +  
 labs(title = "Relationship between Weight and Height by diabates status",  
 x = "Weight", y = "Height",  
 color = "Diabetes") +  
 theme\_minimal()



# Scatter plot with points colored by smoking status  
ggplot(nhanes\_replaced, aes(x = Weight, y = Height, color = SmokingStatus)) +  
 geom\_point(size = 3) +  
 labs(title = "Relationship between Weight and Height by smoking status",  
 x = "Weight", y = "Height",  
 color = "Smoking status") +  
 theme\_minimal()



#Independent samples t-test for age and gender  
t.test(Age ~ Gender, data = nhanes\_data)

##   
## Welch Two Sample t-test  
##   
## data: Age by Gender  
## t = 1.7498, df = 4992.5, p-value = 0.08022  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## -0.1344235 2.3672964  
## sample estimates:  
## mean in group female mean in group male   
## 37.26733 36.15090

#Independent samples t-test for BMI,166 and Diabetes,64  
t.test(BMI ~ Diabetes, data = nhanes\_data)

##   
## Welch Two Sample t-test  
##   
## data: BMI by Diabetes  
## t = -15.907, df = 416.81, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0  
## 95 percent confidence interval:  
## -7.398449 -5.771044  
## sample estimates:  
## mean in group No mean in group Yes   
## 25.94804 32.53279

#Independent samples t-test for Alcohol Year,2016 and Relationship Status,1415  
t.test(AlcoholYear ~ RelationshipStatus, data = nhanes\_data)

##   
## Welch Two Sample t-test  
##   
## data: AlcoholYear by RelationshipStatus  
## t = 5.4315, df = 2674.8, p-value = 6.09e-08  
## alternative hypothesis: true difference in means between group Committed and group Single is not equal to 0  
## 95 percent confidence interval:  
## 13.05949 27.81603  
## sample estimates:  
## mean in group Committed mean in group Single   
## 83.93416 63.49640