Project: Public Health: National Nutritional Health

Goodness Nwokebu

2024-01-29

Introduction

NHANES is a program run by the CDC to assess the health and nutritional status of adults and children in the US. It combines survey questions and physical examinations, including medical and physiological measurements and laboratory tests, and examines a representative sample of about 5,000 people each year. The data is used to determine the prevalence of diseases and risk factors, establish national standards, and support epidemiology studies and health sciences research. This information helps to develop public health policy, design health programs and services, and expand the nation's health knowledge.

Link to Dataset: https://raw.githubusercontent.com/HackBio-Internship/public_datasets/main/R/nhanes.csv

TASK 1: Process all NA (either by deleting or by converting to zero)

```
knitr::opts_chunk$set(echo = TRUE)
#Importing Necessary Libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
#Importing Data and reading the data into a variable
url <- "https://raw.githubusercontent.com/HackBio-Internship/public_datasets/main/R/nhanes.csv"
Data <- read.table(url,header = TRUE,sep = ",")

cat("Number of rows:", nrow(Data), "\n")
```

```
## Number of rows: 5000
```

```r
cat("Number of columns:", ncol(Data), "\n")
```

```
## Number of columns: 32
```

```r
#getting the summary statistics of the data
summary(Data)
```

```
##        id            Gender                Age             Race          
##  Min.   :62163   Length:5000        Min.   : 0.00   Length:5000       
##  1st Qu.:64544   Class :character   1st Qu.:17.00   Class :character  
##  Median :67039   Mode  :character   Median :36.00   Mode  :character  
##  Mean   :67028                      Mean   :36.71                     
##  3rd Qu.:69509                      3rd Qu.:54.00                     
##  Max.   :71915                      Max.   :80.00                     
##                                                                       
##   Education         MaritalStatus      RelationshipStatus   Insured         
##  Length:5000        Length:5000        Length:5000        Length:5000       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##      Income          Poverty        HomeRooms        HomeOwn         
##  Min.   :  2500   Min.   :0.000   Min.   : 1.000   Length:5000       
##  1st Qu.: 30000   1st Qu.:1.190   1st Qu.: 4.000   Class :character  
##  Median : 50000   Median :2.600   Median : 6.000   Mode  :character  
##  Mean   : 57078   Mean   :2.761   Mean   : 6.193                     
##  3rd Qu.:100000   3rd Qu.:4.760   3rd Qu.: 8.000                     
##  Max.   :100000   Max.   :5.000   Max.   :13.000                     
##  NA's   :377      NA's   :325     NA's   :28                         
##      Work               Weight           Height           BMI        
##  Length:5000        Min.   :  3.60   Min.   : 83.8   Min.   :12.90  
##  Class :character   1st Qu.: 55.40   1st Qu.:156.5   1st Qu.:21.50  
##  Mode  :character   Median : 72.10   Median :165.6   Median :25.80  
##                     Mean   : 70.33   Mean   :161.5   Mean   :26.44  
##                     3rd Qu.: 88.10   3rd Qu.:174.2   3rd Qu.:30.60  
##                     Max.   :198.70   Max.   :200.4   Max.   :80.60  
##                     NA's   :31       NA's   :159     NA's   :166    
##      Pulse            BPSys            BPDia        Testosterone      
##  Min.   : 40.00   Min.   : 79.0   Min.   :  0.0   Min.   :   0.25  
##  1st Qu.: 66.00   1st Qu.:107.0   1st Qu.: 62.0   1st Qu.:  17.70  
##  Median : 72.00   Median :116.0   Median : 69.0   Median :  43.82  
##  Mean   : 73.63   Mean   :118.7   Mean   : 68.3   Mean   : 197.90  
##  3rd Qu.: 82.00   3rd Qu.:128.0   3rd Qu.: 77.0   3rd Qu.: 362.41  
##  Max.   :136.00   Max.   :221.0   Max.   :116.0   Max.   :1795.60  
##  NA's   :718      NA's   :719     NA's   :719     NA's   :874      
##     HDLChol          TotChol         Diabetes          DiabetesAge   
##  Min.   :0.410   Min.   : 1.530   Length:5000        Min.   : 1.00  
##  1st Qu.:1.090   1st Qu.: 4.060   Class :character   1st Qu.:39.00  
##  Median :1.290   Median : 4.730   Mode  :character   Median :50.00  
##  Mean   :1.361   Mean   : 4.831                      Mean   :47.61  
##  3rd Qu.:1.580   3rd Qu.: 5.510                      3rd Qu.:57.00  
```

```
##  Max.   :4.030   Max.   :12.280                     Max.   :80.00
##  NA's   :775     NA's   :775                        NA's   :4693
##   nPregnancies       nBabies       SleepHrsNight     PhysActive
##  Min.   : 1.000   Min.   : 0.000   Min.   : 2.000   Length:5000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 6.000   Class :character
##  Median : 3.000   Median : 2.000   Median : 7.000   Mode  :character
##  Mean   : 2.924   Mean   : 2.375   Mean   : 6.906
##  3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.: 8.000
##  Max.   :13.000   Max.   :11.000   Max.   :12.000
##  NA's   :3735     NA's   :3832     NA's   :1166
##  PhysActiveDays    AlcoholDay      AlcoholYear     SmokingStatus
##  Min.   :1.000   Min.   : 1.000   Min.   :  0.00   Length:5000
##  1st Qu.:2.000   1st Qu.: 1.000   1st Qu.:  3.00   Class :character
##  Median :4.000   Median : 2.000   Median : 24.00   Mode  :character
##  Mean   :3.819   Mean   : 2.925   Mean   : 74.86
##  3rd Qu.:5.000   3rd Qu.: 3.000   3rd Qu.:104.00
##  Max.   :7.000   Max.   :82.000   Max.   :364.00
##  NA's   :2614    NA's   :2503     NA's   :2016
```

```
Data_tb = tibble(Data)
```

Data Cleaning

```
#Diabetes and DiabetesAge Column
fd_tb <- Data_tb[
  complete.cases(Data_tb$Diabetes), #delete the nulls from diabetes column
  ]
fd_tb %>% count(
  Diabetes, DiabetesAge) #check if the nullset in DiabetesAge aligns with No of Diabetes column
```

```
## # A tibble: 61 x 3
##    Diabetes DiabetesAge     n
##    <chr>          <int> <int>
##  1 No                NA  4563
##  2 Yes                1     8
##  3 Yes                3     1
##  4 Yes                8     1
##  5 Yes               11     2
##  6 Yes               15     2
##  7 Yes               16     2
##  8 Yes               17     2
##  9 Yes               18     2
## 10 Yes               20     1
## # i 51 more rows
```

```
fd_tb$DiabetesAge[(fd_tb$Diabetes == "No")] <- 0 #equating those with no record of diabetes with zero d
fd_tb <- fd_tb[complete.cases(fd_tb$DiabetesAge),] #delete the remaining nulls from DiabetesAge column

cat("Null value in Diabetes Column:", sum(is.na(fd_tb$Diabetes)), "\n")
```

```
## Null value in Diabetes Column: 0
```

```r
cat ("null value in DiabetesAge Column:", sum(is.na(fd_tb$DiabetesAge)), "\n")
```

```
## null value in DiabetesAge Column: 0
```

```r
#npregnanies Column
fd_tb$nPregnancies[is.na(fd_tb$nPregnancies)] <- 0 #equating null to having zero pregnancies.

#BMI, PULSE, BPSys, BPDia, Testerone
col_to_del = c("BMI", "Pulse", "BPSys", 'BPDia', "Testosterone", 'HDLChol','TotChol') #delete null col
filteredata_tb <- subset(fd_tb, complete.cases(fd_tb[, col_to_del]))

#nBabies
nBabies_NA <- subset(filteredata_tb, is.na(nBabies))
filteredata_tb %>% count(nPregnancies, nBabies) #getting the summary data of nPrgnancies and nBabies.
```

```
## # A tibble: 40 x 3
##    nPregnancies nBabies     n
##           <dbl>   <int> <int>
##  1            0      NA  2712
##  2            1       1   141
##  3            1      NA    67
##  4            2       0     2
##  5            2       1    62
##  6            2       2   247
##  7            2      NA     9
##  8            3       1    40
##  9            3       2    92
## 10            3       3   131
## # i 30 more rows
```

```r
#filling null values in nBabies where pregnancy is 0 with 0.
filteredata_tb <- filteredata_tb %>% mutate( nBabies_filled = ifelse(nPregnancies ==0 & is.na(nBabies),
filt_tb <- filteredata_tb %>% group_by( nPregnancies) %>% mutate(
  nBabies_filled = ifelse(is.na(nBabies_filled), round(mean(nBabies_filled, na.rm = TRUE), digits = 0),
filteredata_tb <- filteredata_tb %>% mutate(
  nBabies_filled = filt_tb$nBabies_filled)

filteredata_tb$nBabies = NULL  #removing the nBabies column

#Alcoholday, AlcoholYear, Income, HomeRooms, SmokingStatus
filteredata_tb$AlcoholYear[is.na(filteredata_tb$AlcoholYear)] <- 0
filteredata_tb$AlcoholDay[is.na(filteredata_tb$AlcoholDay)] <- 0
filteredata_tb$Income[is.na(filteredata_tb$Income)] <- 0
filteredata_tb$HomeRooms[is.na(filteredata_tb$HomeRooms)] <-0
filteredata_tb$SmokingStatus[is.na(filteredata_tb$SmokingStatus)] <- "Never"
```

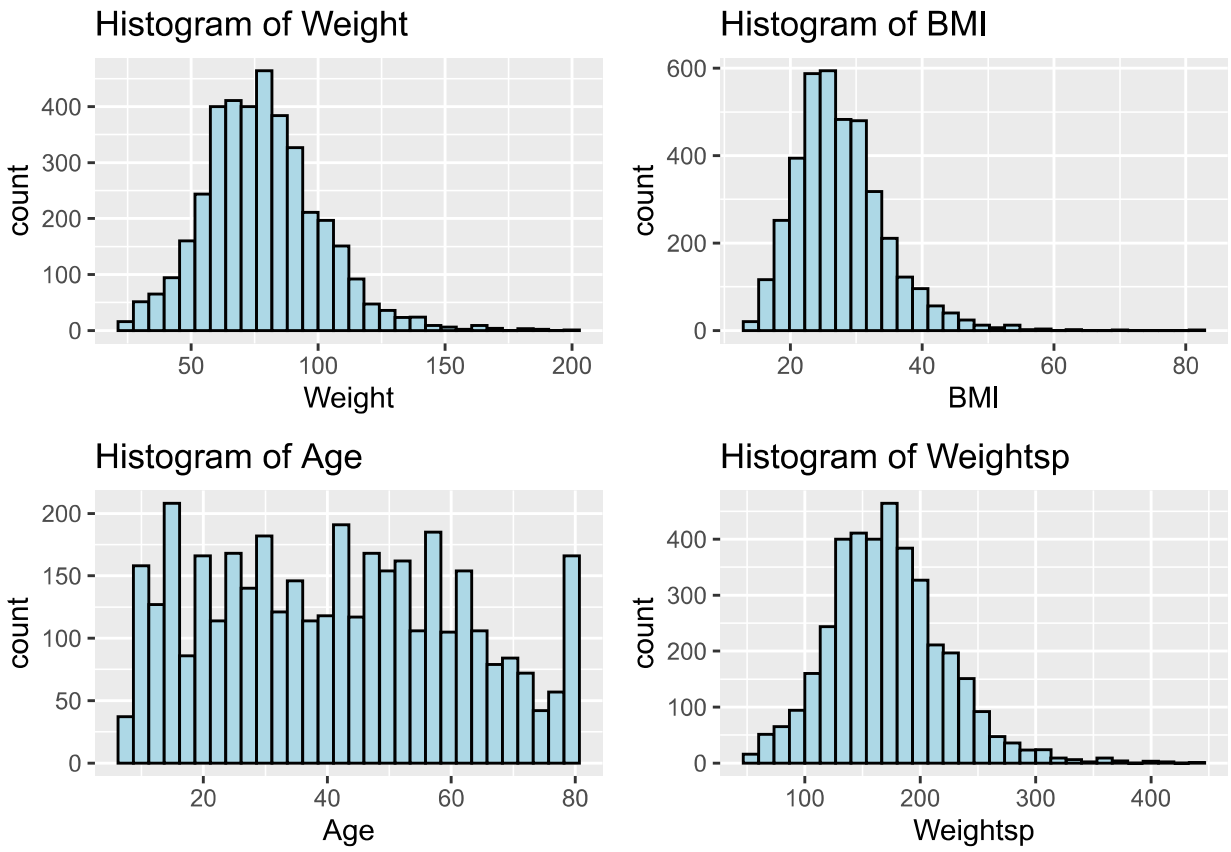
TASK 2: Visualize the distribution of BMI, Weight, Weight in pounds (weight *2.2) and Age with an histogram.

```r
#Data Analysis
dataplot1 <- select(filteredata_tb,Weight, BMI, Age)
#computing weight in pounds
dataplot1 <- dataplot1 %>% mutate(Weightsp = Weight * 2.2)

#plotting thr histograms
plots <- list()

for (col_name in names(dataplot1)) {
  # Use .. notation to dynamically refer to the column in aes()
  plot <- ggplot(dataplot1, aes_string(x = col_name)) +
    geom_histogram(fill = "lightblue", col = 'black') +
    labs(title = paste("Histogram of", col_name))

  plots[[col_name]] <- plot
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Arrange and print the plots
gridExtra::grid.arrange(grobs = plots, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The four histogram shows the distribution of Weight, BMI, Age and the Weight(in pounds). Excerpts: 1) Most of the population's weight fell between 70 - 80kg and 140-180 pounds 2) Participants aged 60 and above were less represented in the population 3) A lot of the participants seems to be obese having BMI >30

```
round(mean(filteredata_tb$Pulse)) #mean pulse of the population
```

TASK 3: What's the mean 60-second pulse rate for all participants in the data?

```
## [1] 74
```

The population mean pulse is 74

```
paste("The range of diastolic blood pressure of all participant is",
      min(filteredata_tb$BPDia), "-", max(filteredata_tb$BPDia))
```

TASK 4: What's the range of values for diastolic blood pressure in all participants?

```
## [1] "The range of diastolic blood pressure of all participant is 0 - 116"
```

```r
paste("The standard deviation and variance of the partcipants income  are",
      round(sd(filteredata_tb$Income)), "and", round(var(filteredata_tb$Income)),
      "respectively")
```

TASK 5: What's the variance and standard deviation for income among all participants?

```
## [1] "The standard deviation and variance of the partcipants income  are 35528 and 1262267482 respect
```

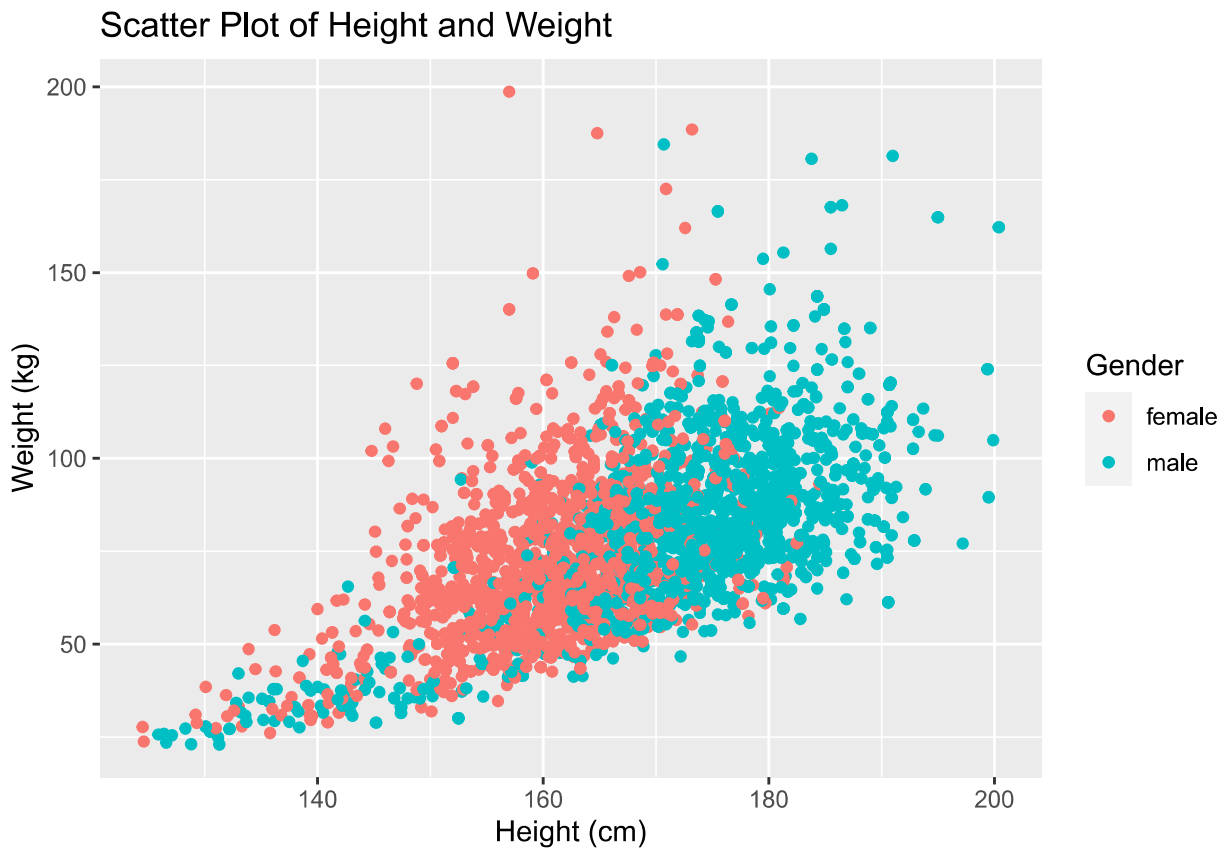
This shows that there is a significant difference in the individual's income. This maybe from the fact that about 20% of the population are under eighteen years with a no official means of income.

TASK 6: Visualize the relationship between weight and height

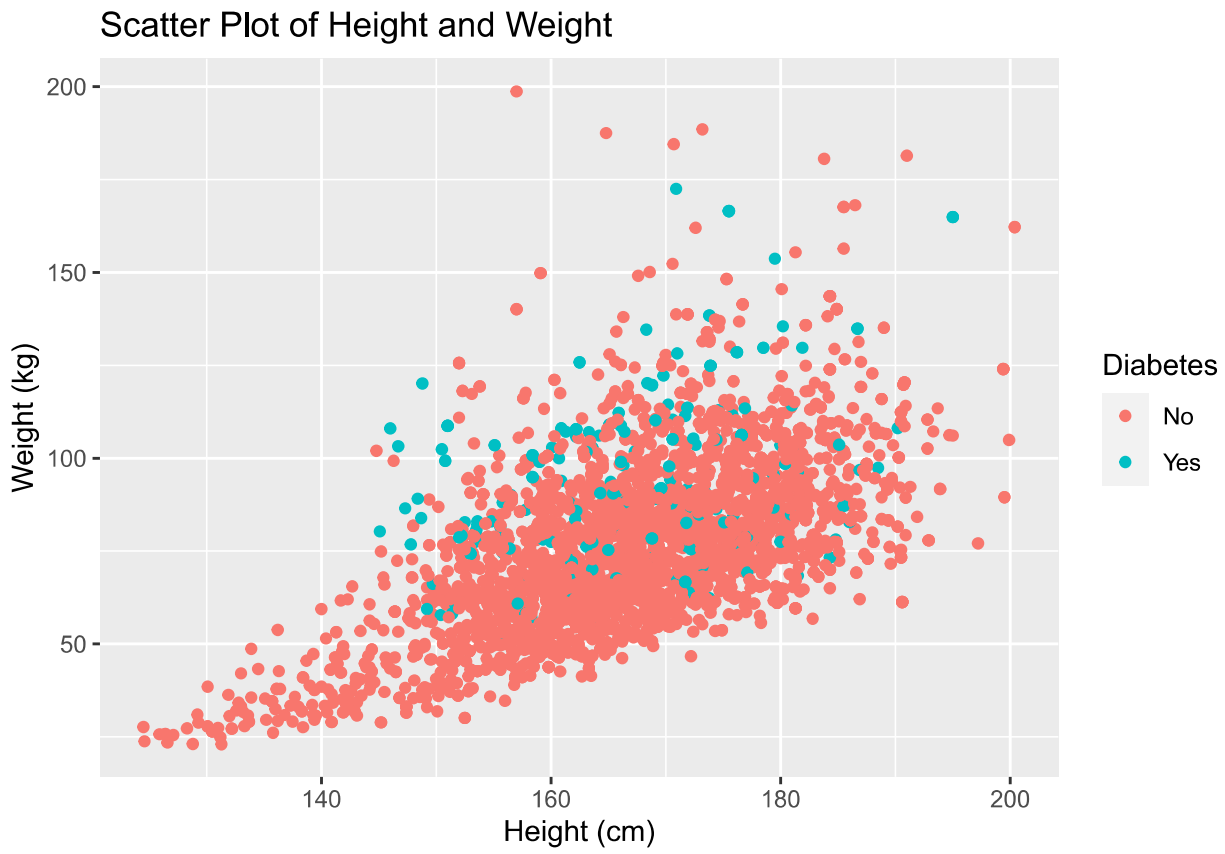
Color the points by:

- gender
- diabetes
- smoking status

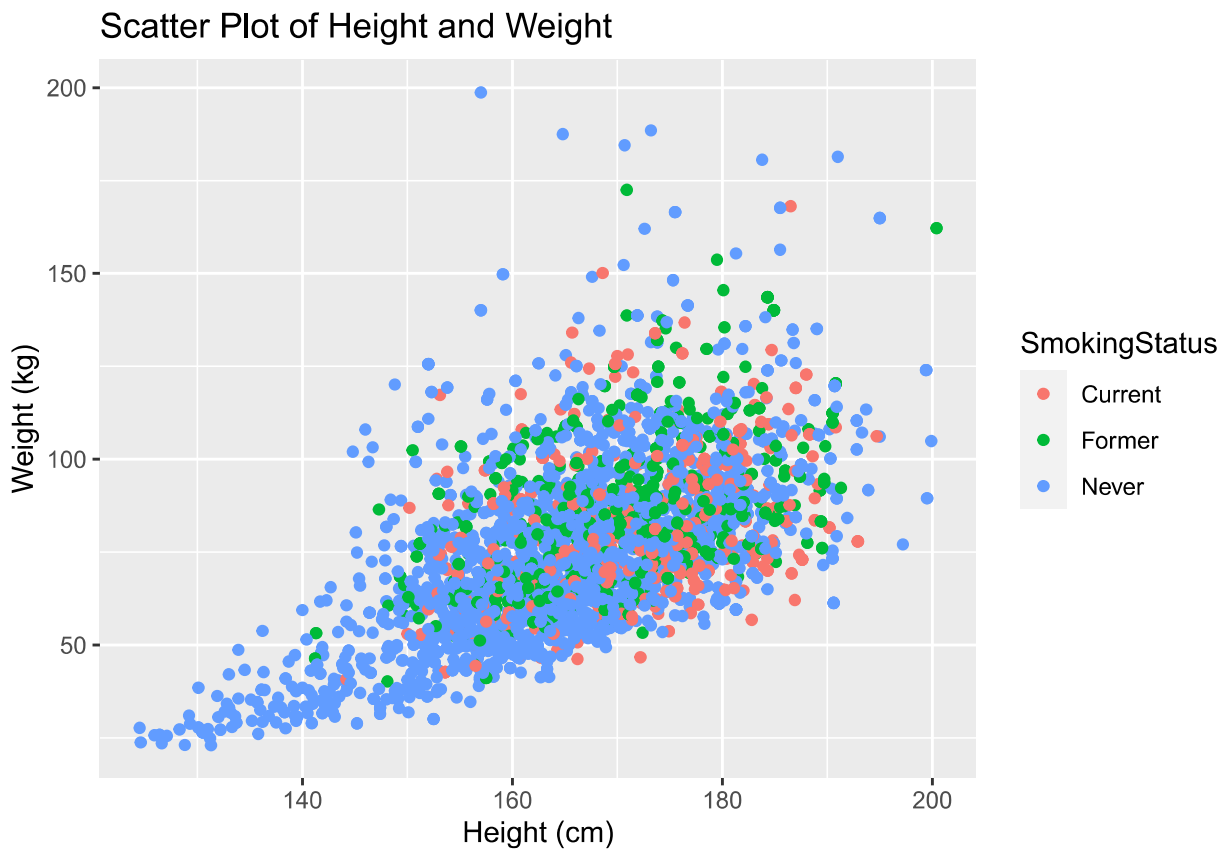
```r
#Colour by Gender
ggplot(filteredata_tb, aes(x = Height,y = Weight,
                           color = Gender))+ geom_point() + labs(
                             title = "Scatter Plot of Height and Weight",
                           x = "Height (cm)", y = "Weight (kg)")
```



```
#colour by diabetes
ggplot(filteredata_tb, aes(x = Height, y = Weight,
                           color = Diabetes)) + geom_point() + labs(
                             title = "Scatter Plot of Height and Weight",
                           x = "Height (cm)", y = "Weight (kg)")
```




```
#Colour by Smoking Status
ggplot(filteredata_tb, aes(x = Height, y = Weight,
                           color = SmokingStatus)) + geom_point() + labs(
                             title = "Scatter Plot of Height and Weight",
                             x = "Height (cm)", y = "Weight (kg)")
```



TASK 7: Conduct t-test between the following variables and make conclusions on the relationship between them based on P-Value

- Age and Gender
- BMI and Diabetes
- Alcohol Year and Relationship Status

```
print(t.test(Age ~ Gender, data= filteredata_tb))
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Gender
## t = 0.93329, df = 3813.1, p-value = 0.3507
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -0.6649573  1.8731817
## sample estimates:
## mean in group female   mean in group male
##             41.71557             41.11146
```

This test aims to check if the there is a significant statistical differences between the Female mean age and that of the males. From the result above, female group have an average mean of 41.72 while that of male is 41.11. A positive t-test was gotten meaning there is a positive which fully supports that the mean age of

females is greater than male but it is not statistically significant(P-value > 0.05). This means that we do not have enough evidence to reject the null hypothesis. Hence the variability in the means can be random either than a true difference.

```
print(t.test(BMI ~ Diabetes, data= filteredata_tb))
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by Diabetes
## t = -12.629, df = 298.91, 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:
##  -6.941702 -5.070018
## sample estimates:
##  mean in group No mean in group Yes
##          27.15957          33.16543
```

The mean BMI of people with Diabetes(Diabetes is equal to Yes) is higher than those without(Diabetes is equal to No). P_value of 2.2E-16 is lesser than 0.05 and the confidence interval not including 0 shows that this difference is statistically significant and likely not by chance.

```
print(t.test(AlcoholYear~RelationshipStatus, data= filteredata_tb))
```

```
##
##  Welch Two Sample t-test
##
## data:  AlcoholYear by RelationshipStatus
## t = 4.808, df = 2950.4, p-value = 1.601e-06
## alternative hypothesis: true difference in means between group Committed and group Single is not equ
## 95 percent confidence interval:
##   9.703328 23.067929
## sample estimates:
## mean in group Committed    mean in group Single
##                69.41839                53.03276
```

The small p-value and the confidence interval not including 0 suggest strong evidence that the mean alcohol consumption is different between individuals in the "Committed" and "Single" relationship statuses. The positive t-value indicates that, on average, individuals in the "Committed" group have a higher mean alcohol consumption than those in the "Single" group. The difference in means is estimated to be between 9.70 and 23.07 units.