Week 8 A

Wali Reheman

2024-10-09

Dataset

We will use the **NHANES** dataset, which contains health and demographic information collected from a representative sample of the U.S. population. The dataset is available in R through the `NHANES` package.

Accessing the Dataset:

```
#install.packages("NHANES")  # Install if not already installed
library(NHANES)
data <- data.frame(NHANES)
```

TASK 1: Data Exploration

- **Load the `NHANES` dataset**.
- Use `tidyverse` functions to:
  - View the **structure** of the dataset.
  - Summarize key variables: `Age`, `Gender`, `BMI`, `SmokeNow` (whether the person currently smokes), and `PhysActive`(physical activity status).
  - Just check the **levels** if it is categorical variable.

```
library(tidyverse)

glimpse(data)
```

```
## Rows: 10,000
## Columns: 76
## $ ID              <int> 51624, 51624, 51624, 51625, 51630, 51638, 51646, 5164~
## $ SurveyYr        <fct> 2009_10, 2009_10, 2009_10, 2009_10, 2009_10, 2009_10,~
## $ Gender          <fct> male, male, male, male, female, male, male, female, f~
## $ Age             <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 10, ~
## $ AgeDecade       <fct>  30-39,  30-39,  30-39,  0-9,  40-49,  0-9,  0-9,  40~
## $ AgeMonths       <int> 409, 409, 409, 49, 596, 115, 101, 541, 541, 541, 795,~
## $ Race1           <fct> White, White, White, Other, White, White, White, Whit~
## $ Race3           <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ Education       <fct> High School, High School, High School, NA, Some Colle~
## $ MaritalStatus   <fct> Married, Married, Married, NA, LivePartner, NA, NA, M~
## $ HHIncome        <fct> 25000-34999, 25000-34999, 25000-34999, 20000-24999, 3~
## $ HHIncomeMid     <int> 30000, 30000, 30000, 22500, 40000, 87500, 60000, 8750~
```

```
## $ Poverty          <dbl> 1.36, 1.36, 1.36, 1.07, 1.91, 1.84, 2.33, 5.00, 5.00,~
## $ HomeRooms        <int> 6, 6, 6, 9, 5, 6, 7, 6, 6, 6, 5, 10, 6, 10, 10, 4, 3,~
## $ HomeOwn          <fct> Own, Own, Own, Own, Rent, Rent, Own, Own, Own, Own, O~
## $ Work             <fct> NotWorking, NotWorking, NotWorking, NA, NotWorking, N~
## $ Weight           <dbl> 87.4, 87.4, 87.4, 17.0, 86.7, 29.8, 35.2, 75.7, 75.7,~
## $ Length           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ HeadCirc         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ Height           <dbl> 164.7, 164.7, 164.7, 105.4, 168.4, 133.1, 130.6, 166.~
## $ BMI              <dbl> 32.22, 32.22, 32.22, 15.30, 30.57, 16.82, 20.64, 27.2~
## $ BMICatUnder20yrs <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ BMI_WHO          <fct> 30.0_plus, 30.0_plus, 30.0_plus, 12.0_18.5, 30.0_plus~
## $ Pulse            <int> 70, 70, 70, NA, 86, 82, 72, 62, 62, 62, 60, 62, 76, 8~
## $ BPSysAve         <int> 113, 113, 113, NA, 112, 86, 107, 118, 118, 118, 111, ~
## $ BPDiaAve         <int> 85, 85, 85, NA, 75, 47, 37, 64, 64, 64, 63, 74, 85, 6~
## $ BPSys1           <int> 114, 114, 114, NA, 118, 84, 114, 106, 106, 106, 124, ~
## $ BPDia1           <int> 88, 88, 88, NA, 82, 50, 46, 62, 62, 62, 64, 76, 86, 6~
## $ BPSys2           <int> 114, 114, 114, NA, 108, 84, 108, 118, 118, 118, 108, ~
## $ BPDia2           <int> 88, 88, 88, NA, 74, 50, 36, 68, 68, 68, 62, 72, 88, 6~
## $ BPSys3           <int> 112, 112, 112, NA, 116, 88, 106, 118, 118, 118, 114, ~
## $ BPDia3           <int> 82, 82, 82, NA, 76, 44, 38, 60, 60, 60, 64, 76, 82, 7~
## $ Testosterone     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ DirectChol       <dbl> 1.29, 1.29, 1.29, NA, 1.16, 1.34, 1.55, 2.12, 2.12, 2~
## $ TotChol          <dbl> 3.49, 3.49, 3.49, NA, 6.70, 4.86, 4.09, 5.82, 5.82, 5~
## $ UrineVol1        <int> 352, 352, 352, NA, 77, 123, 238, 106, 106, 106, 113, ~
## $ UrineFlow1       <dbl> NA, NA, NA, NA, 0.094, 1.538, 1.322, 1.116, 1.116, 1.~
## $ UrineVol2        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ UrineFlow2       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ Diabetes         <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, N~
## $ DiabetesAge      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ HealthGen        <fct> Good, Good, Good, NA, Good, NA, NA, Vgood, Vgood, Vgo~
## $ DaysPhysHlthBad  <int> 0, 0, 0, NA, 0, NA, NA, 0, 0, 0, 10, 0, 4, NA, NA, 0,~
## $ DaysMentHlthBad  <int> 15, 15, 15, NA, 10, NA, NA, 3, 3, 3, 0, 0, 0, NA, NA,~
## $ LittleInterest   <fct> Most, Most, Most, NA, Several, NA, NA, None, None, No~
## $ Depressed        <fct> Several, Several, Several, NA, Several, NA, NA, None,~
## $ nPregnancies     <int> NA, NA, NA, NA, 2, NA, NA, 1, 1, 1, NA, NA, NA, NA, N~
## $ nBabies          <int> NA, NA, NA, NA, 2, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ Age1stBaby       <int> NA, NA, NA, NA, 27, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ SleepHrsNight    <int> 4, 4, 4, NA, 8, NA, NA, 8, 8, 8, 7, 5, 4, NA, 5, 7, N~
## $ SleepTrouble     <fct> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, No, No, Y~
## $ PhysActive       <fct> No, No, No, NA, No, NA, NA, Yes, Yes, Yes, Yes, Yes, ~
## $ PhysActiveDays   <int> NA, NA, NA, NA, NA, NA, NA, 5, 5, 5, 7, 5, 1, NA, 2, ~
## $ TVHrsDay         <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ CompHrsDay       <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ TVHrsDayChild    <int> NA, NA, NA, 4, NA, 5, 1, NA, NA, NA, NA, NA, NA, 4, N~
## $ CompHrsDayChild  <int> NA, NA, NA, 1, NA, 0, 6, NA, NA, NA, NA, NA, NA, 3, N~
## $ Alcohol12PlusYr  <fct> Yes, Yes, Yes, NA, Yes, NA, NA, Yes, Yes, Yes, Yes, Y~
## $ AlcoholDay       <int> NA, NA, NA, NA, 2, NA, NA, 3, 3, 3, 1, 2, 6, NA, NA, ~
## $ AlcoholYear      <int> 0, 0, 0, NA, 20, NA, NA, 52, 52, 52, 100, 104, 364, N~
## $ SmokeNow         <fct> No, No, No, NA, Yes, NA, NA, NA, NA, NA, No, NA, NA, ~
## $ Smoke100         <fct> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, Yes, No, ~
## $ Smoke100n        <fct> Smoker, Smoker, Smoker, NA, Smoker, NA, NA, Non-Smoke~
## $ SmokeAge         <int> 18, 18, 18, NA, 38, NA, NA, NA, NA, NA, 13, NA, NA, N~
## $ Marijuana        <fct> Yes, Yes, Yes, NA, Yes, NA, NA, Yes, Yes, Yes, NA, Ye~
## $ AgeFirstMarij    <int> 17, 17, 17, NA, 18, NA, NA, 13, 13, 13, NA, 19, 15, N~
```

```
## $ RegularMarij     <fct> No, No, No, NA, No, NA, NA, No, No, No, NA, Yes, Yes,~
## $ AgeRegMarij      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20, 15, N~
## $ HardDrugs        <fct> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, No, Yes, ~
## $ SexEver          <fct> Yes, Yes, Yes, NA, Yes, NA, NA, Yes, Yes, Yes, Yes, Y~
## $ SexAge           <int> 16, 16, 16, NA, 12, NA, NA, 13, 13, 13, 17, 22, 12, N~
## $ SexNumPartnLife  <int> 8, 8, 8, NA, 10, NA, NA, 20, 20, 20, 15, 7, 100, NA, ~
## $ SexNumPartYear   <int> 1, 1, 1, NA, 1, NA, NA, 0, 0, 0, NA, 1, 1, NA, NA, 1,~
## $ SameSex          <fct> No, No, No, NA, Yes, NA, NA, Yes, Yes, Yes, No, No, N~
## $ SexOrientation   <fct> Heterosexual, Heterosexual, Heterosexual, NA, Heteros~
## $ PregnantNow      <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
# Summarize Age and BMI
summary(data %>% select(Gender,Age, BMI,SmokeNow,PhysActive))
```

```
##     Gender          Age              BMI         SmokeNow    PhysActive
##  female:5020   Min.   : 0.00   Min.   :12.88   No  :1745   No  :3677
##  male  :4980   1st Qu.:17.00   1st Qu.:21.58   Yes :1466   Yes :4649
##                Median :36.00   Median :25.98   NA's:6789   NA's:1674
##                Mean   :36.74   Mean   :26.66
##                3rd Qu.:54.00   3rd Qu.:30.89
##                Max.   :80.00   Max.   :81.25
##                                NA's   :366
```

```
# Frequency table for Gender
table(data$Gender)
```

```
##
## female   male
##   5020   4980
```

```
# Frequency table for SmokeNow
table(data$SmokeNow)
```

```
##
##   No  Yes
## 1745 1466
```

```
# two ways to check levels

print(head(data$Gender))
```

```
## [1] male   male   male   male   female male
## Levels: female male
```

```
levels(data$Gender)
```

```
## [1] "female" "male"
```

TASK 2: Data Cleaning

- Check for **missing values** in key variables.
- Remove any rows with missing values.
- Recode `SmokeNow` and `PhysActive` to have more interpretable categories if necessary.

```
data_clean <- data %>%
  drop_na(Age, BMI, Gender, SmokeNow, PhysActive)




# (Optionally)

# Optional 1: Recode variables for more information when doing summary

data_clean1 <- data_clean %>%
  mutate(
    SmokeNow = recode(SmokeNow, "Yes" = 1, "No" = 0),
    PhysActive = recode(PhysActive, "Yes" = 1, "No" = 0),
    Gender = recode(Gender, "male" = 1, "female" = 0)
  )

summary(data_clean1 %>% select(Gender,Age, BMI,SmokeNow,PhysActive))
```

```
##      Gender            Age              BMI            SmokeNow
##  Min.   :0.0000   Min.   :20.00   Min.   :15.02   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:35.00   1st Qu.:23.90   1st Qu.:0.0000
##  Median :1.0000   Median :49.00   Median :27.54   Median :0.0000
##  Mean   :0.5625   Mean   :48.91   Mean   :28.45   Mean   :0.4579
##  3rd Qu.:1.0000   3rd Qu.:62.00   3rd Qu.:31.90   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :80.00   Max.   :67.83   Max.   :1.0000
##    PhysActive
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4771
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
# Option 2: Recode variables for regression table later

# this way so that your regrestion table is more straitforward

data_clean2 <- data_clean %>%
  mutate(
    SmokeNow = recode(SmokeNow, "Yes" = "Smoker", "No" = "Non-Smoker"),
    PhysActive = recode(PhysActive, "Yes" = "Active", "No" = "Inactive"))

summary(data_clean2 %>% select(Gender,Age, BMI,SmokeNow,PhysActive))
```

```
##     Gender           Age              BMI              SmokeNow
```

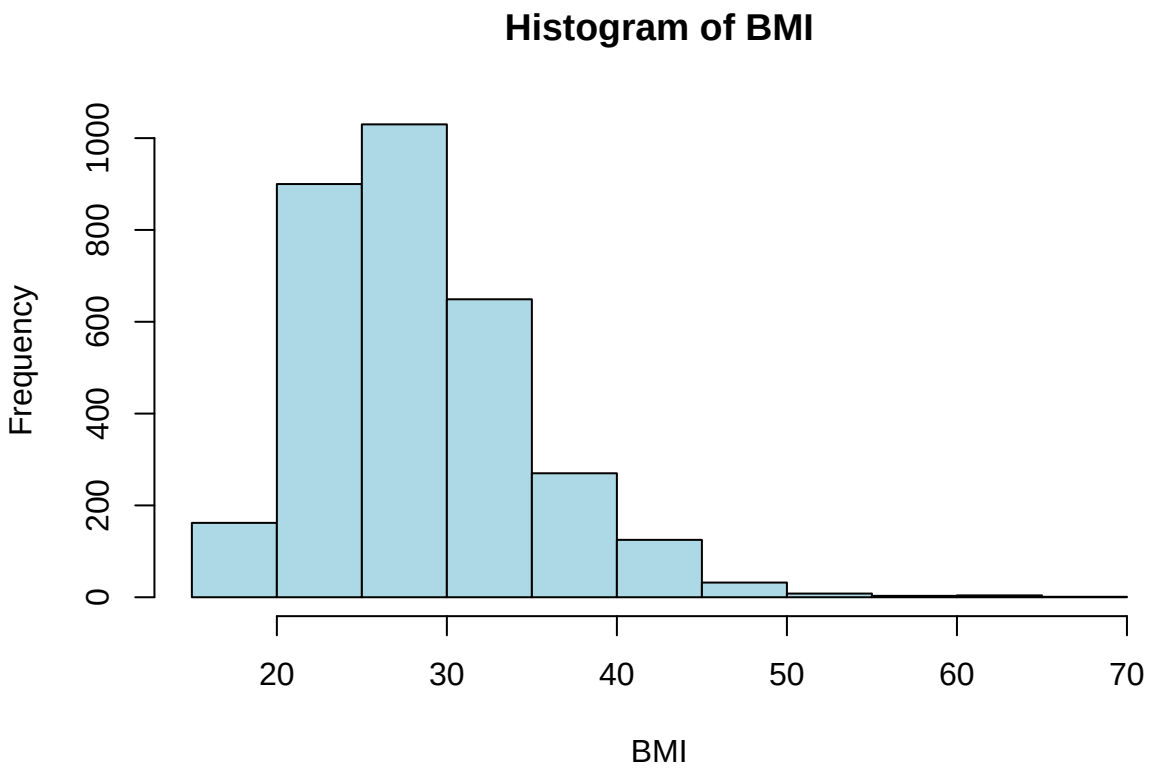
```
##  female:1393   Min.   :20.00   Min.   :15.02   Non-Smoker:1726
##  male  :1791   1st Qu.:35.00   1st Qu.:23.90   Smoker    :1458
##                Median :49.00   Median :27.54
##                Mean   :48.91   Mean   :28.45
##                3rd Qu.:62.00   3rd Qu.:31.90
##                Max.   :80.00   Max.   :67.83
##     PhysActive
##  Inactive:1665
##  Active  :1519
##
##
##
##
```

TASK 3: Data Visualization

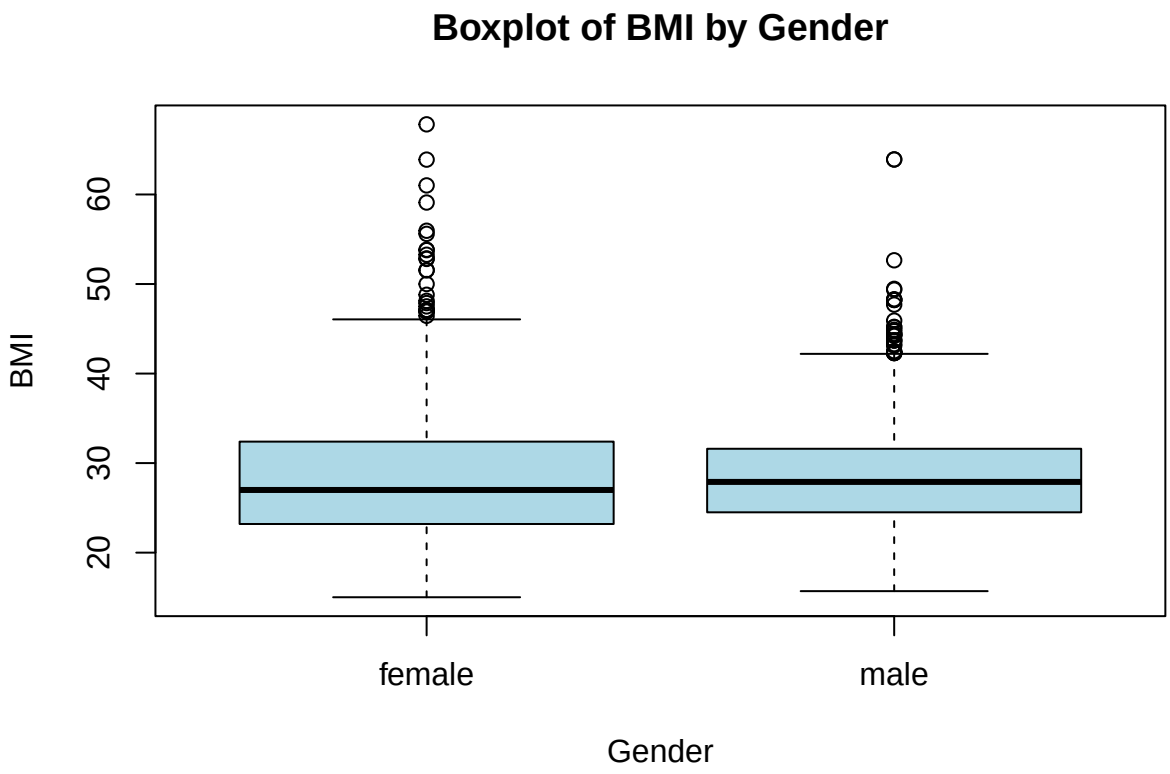
- Create a **histogram** of `BMI`.
- Create a **boxplot** of `BMI` by `Gender`.
- Add appropriate **labels** and **titles**.

```r
## basic R

hist(data_clean2$BMI,
     main = "Histogram of BMI",
     xlab = "BMI",
     col = "lightblue",  # Optional: Adds color to the bars
     border = "black")   # Optional: Adds a border to the bars
```

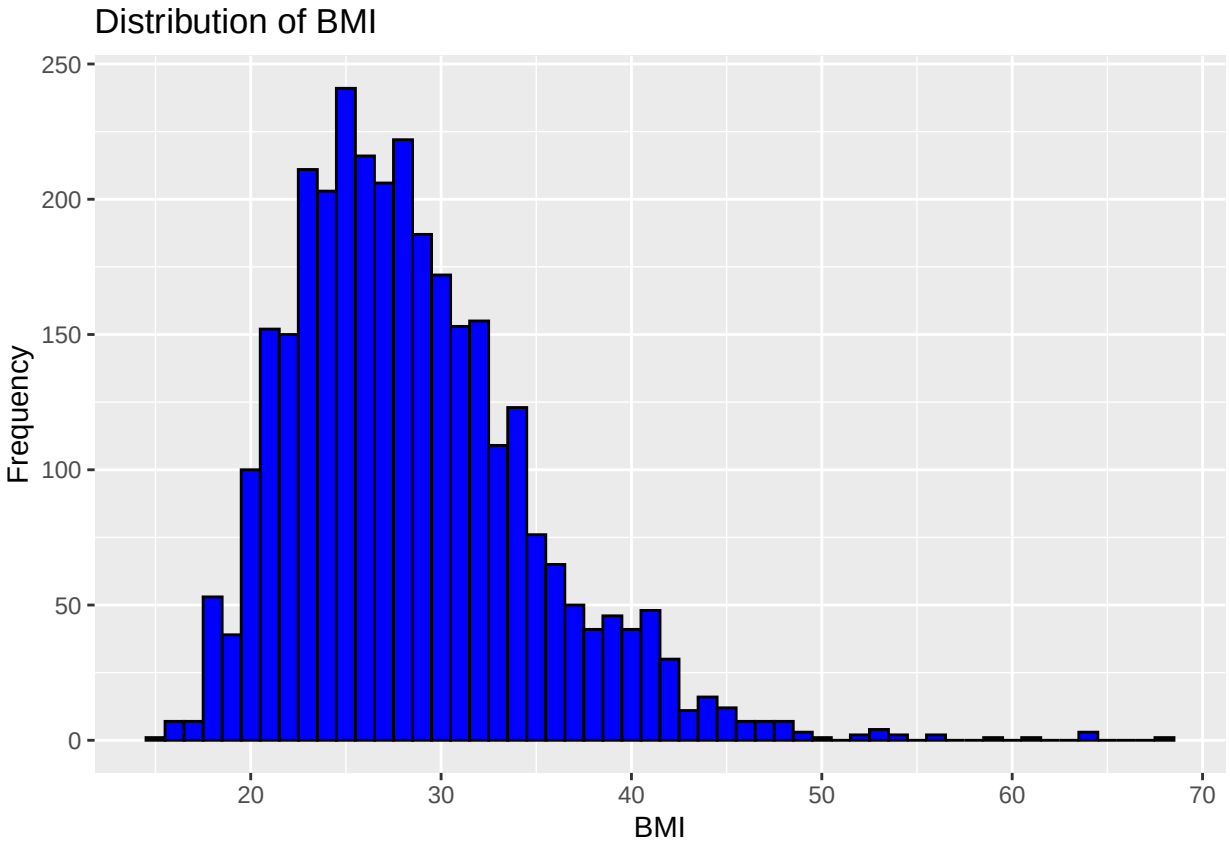



```r
boxplot(BMI ~ Gender, data = data_clean2,  # remember to use the dataset you didn't code Gender as numb
        main = "Boxplot of BMI by Gender",
        xlab = "Gender",
        ylab = "BMI",
        col = "lightblue")  # Optional: Adds color to the boxplot
```

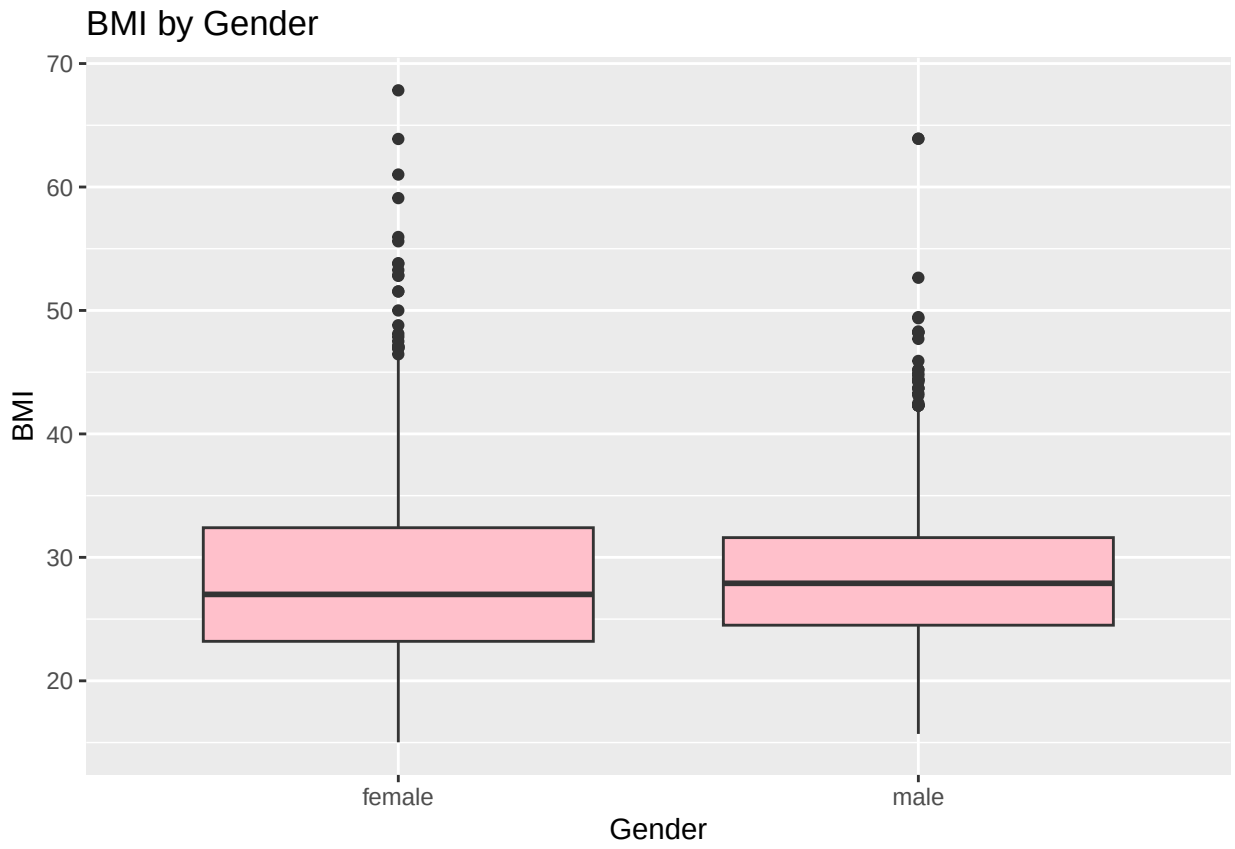



```
# ggplot2 package


ggplot(data_clean1, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +  # we are doing it a bit differently by
  labs(title = "Distribution of BMI", x = "BMI", y = "Frequency")
```




```
ggplot(data_clean2, aes(x = Gender, y = BMI)) +
  geom_boxplot(fill = "pink") +
  labs(title = "BMI by Gender", x = "Gender", y = "BMI")
```



TASK 4: Descriptive Statistics

- Use stargazer Present sample statistics for our key variables, mutate the data if needed (for categorical variables)
- Present the results in a **neat table**.

```
# Only do sample statistic for our key variables
key_variables <- data_clean1 %>%
  select(Age, BMI, Gender, SmokeNow, PhysActive)



library(stargazer)

# Key variables: Age,Gender,BMI,SmokeNow(whether the person currently smokes), and PhysActive(physical
stargazer(
  key_variables,
  type = "text", #comment out this line and de-comment the following two line if you need a doc file
#  type = "html",
#  out = "sample statistics.doc",
  title = "Descriptive Statistics",
  digits = 2, # keeping two digit after "."
  summary.stat = c("mean", "sd", "min", "max","n")
)
```

```
##
## Descriptive Statistics
## ==========================================
## Statistic  Mean  St. Dev.  Min   Max    N
## ------------------------------------------
## Age        48.91  16.79    20    80   3,184
## BMI        28.45   6.33   15.02 67.83 3,184
## Gender     0.56    0.50     0     1   3,184
## SmokeNow   0.46    0.50     0     1   3,184
## PhysActive 0.48    0.50     0     1   3,184
## ------------------------------------------
```

TASK 5: Regression Analysis

- Run a linear regression with `BMI` as the **dependent variable** and `Age`, `Gender`, and `PhysActive` as **independent variables**.
- **Interpret** the coefficients (to yourself or group member).

```
model_basic <- lm(BMI ~ Age + Gender + PhysActive, data = data_clean2)
summary(model_basic)
```

```
##
## Call:
## lm(formula = BMI ~ Age + Gender + PhysActive, data = data_clean2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.092  -4.434  -0.957   3.488  38.686
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      28.536543   0.399157  71.492  < 2e-16 ***
## Age               0.015578   0.006699   2.325   0.0201 *
## Gendermale        0.009406   0.223630   0.042   0.9665
## PhysActiveActive -1.785032   0.225212  -7.926  3.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.258 on 3180 degrees of freedom
## Multiple R-squared:  0.02348,    Adjusted R-squared:  0.02256
## F-statistic: 25.49 on 3 and 3180 DF,  p-value: 2.716e-16
```

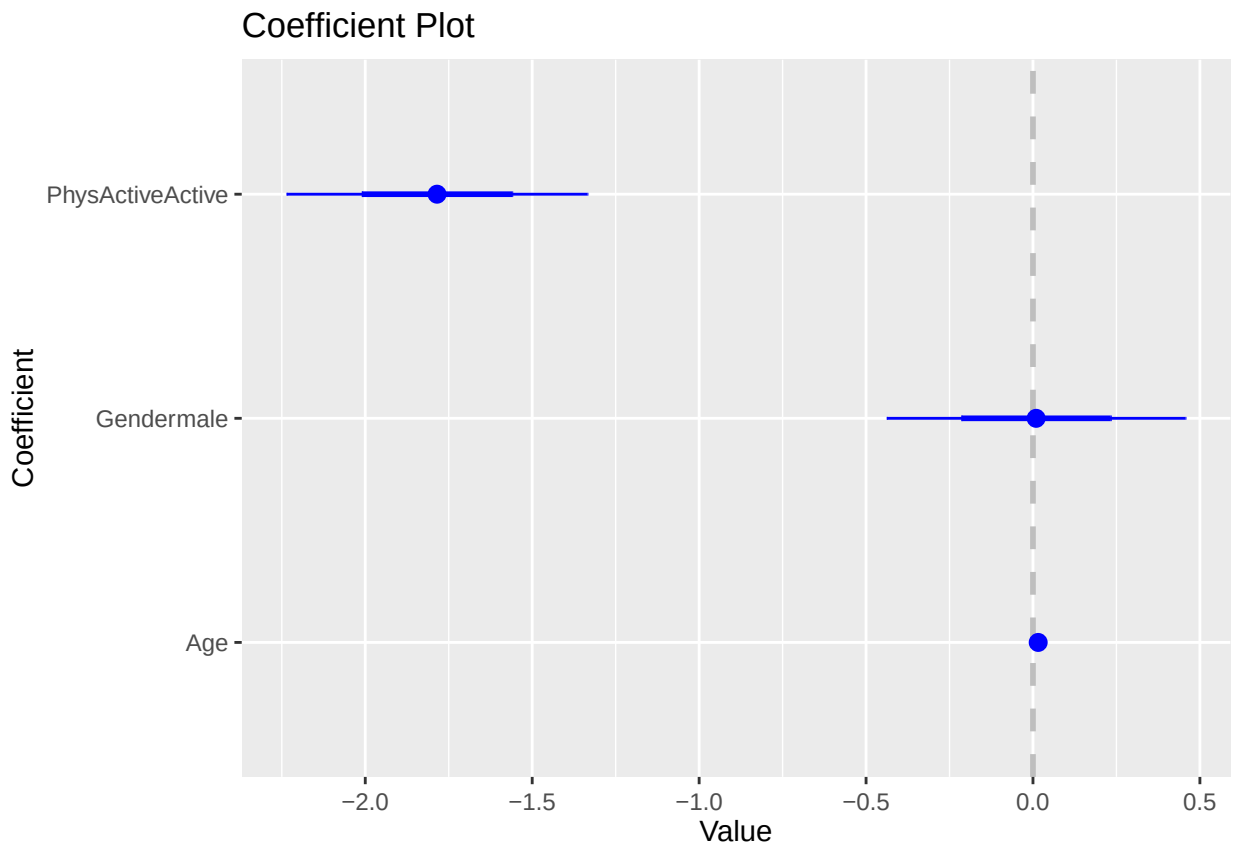
TASK 6: Presenting Results

- Use `stargazer` to create a **regression table (both in the console and a Word doc)**.
- Use `coefplot` to **visualize** the coefficients.

```
stargazer(model_basic, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  BMI
## ------------------------------------------------
## Age                             0.016**
##                                 (0.007)
##
## Gendermale                       0.009
##                                 (0.224)
##
## PhysActiveActive                -1.785***
##                                 (0.225)
##
## Constant                        28.537***
##                                 (0.399)
##
## ------------------------------------------------
## Observations                     3,184
## R2                               0.023
## Adjusted R2                      0.023
## Residual Std. Error       6.258 (df = 3180)
## F Statistic           25.492*** (df = 3; 3180)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```r
library(coefplot)
coefplot(model_basic, title = "Coefficient Plot",intercept = F)
```


Coefficient Plot
PhysActiveActive
Gendermale
Age
Coefficient
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Value