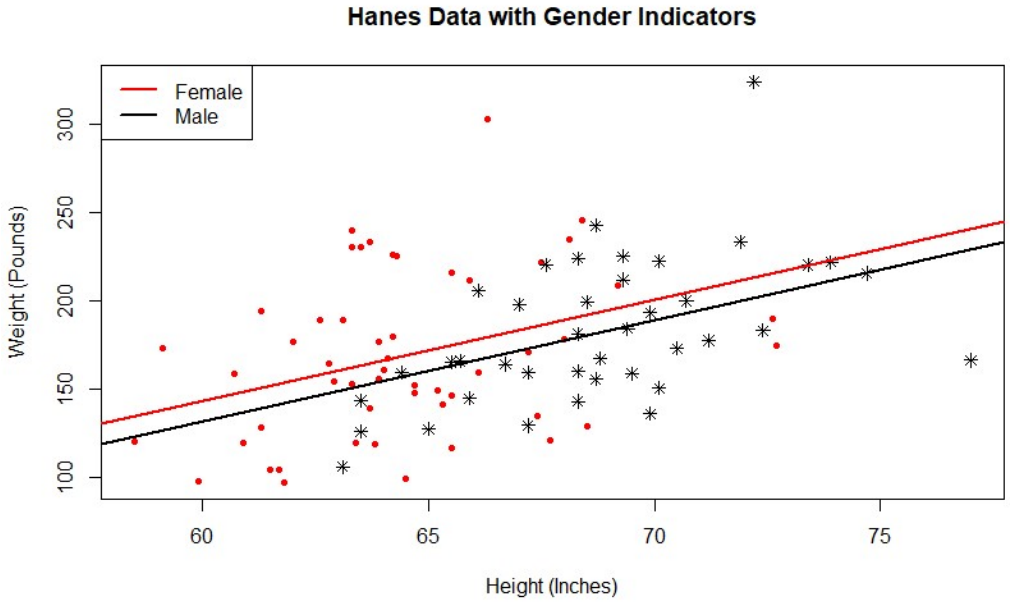
Scatterplot Homework

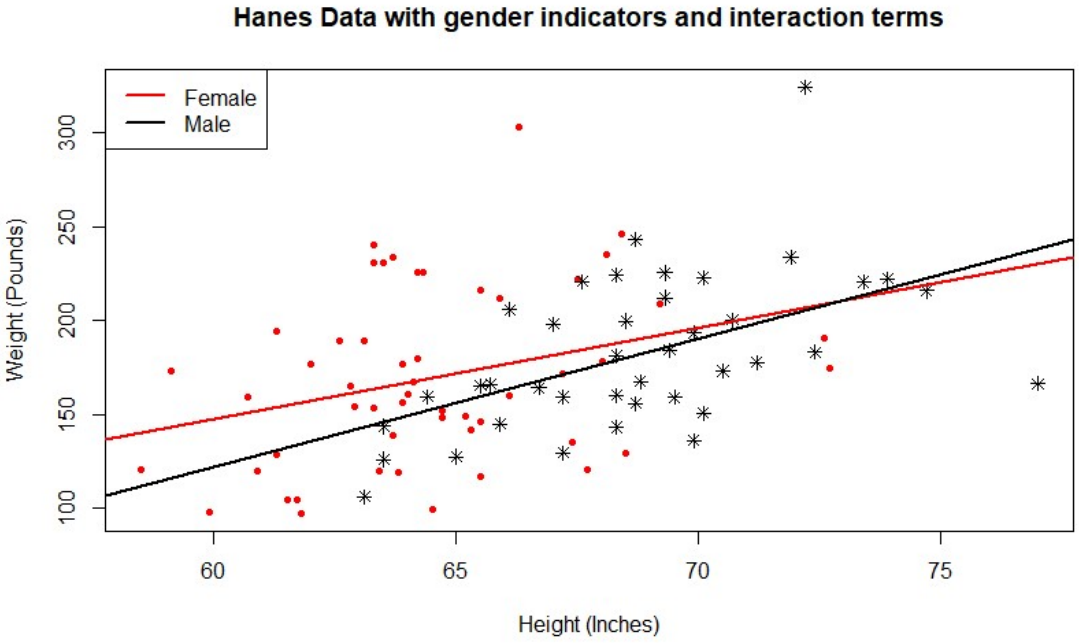
Name: Atharva Shripad Kahu **NetID: axk220142**

Q1 Hanes Data Plot



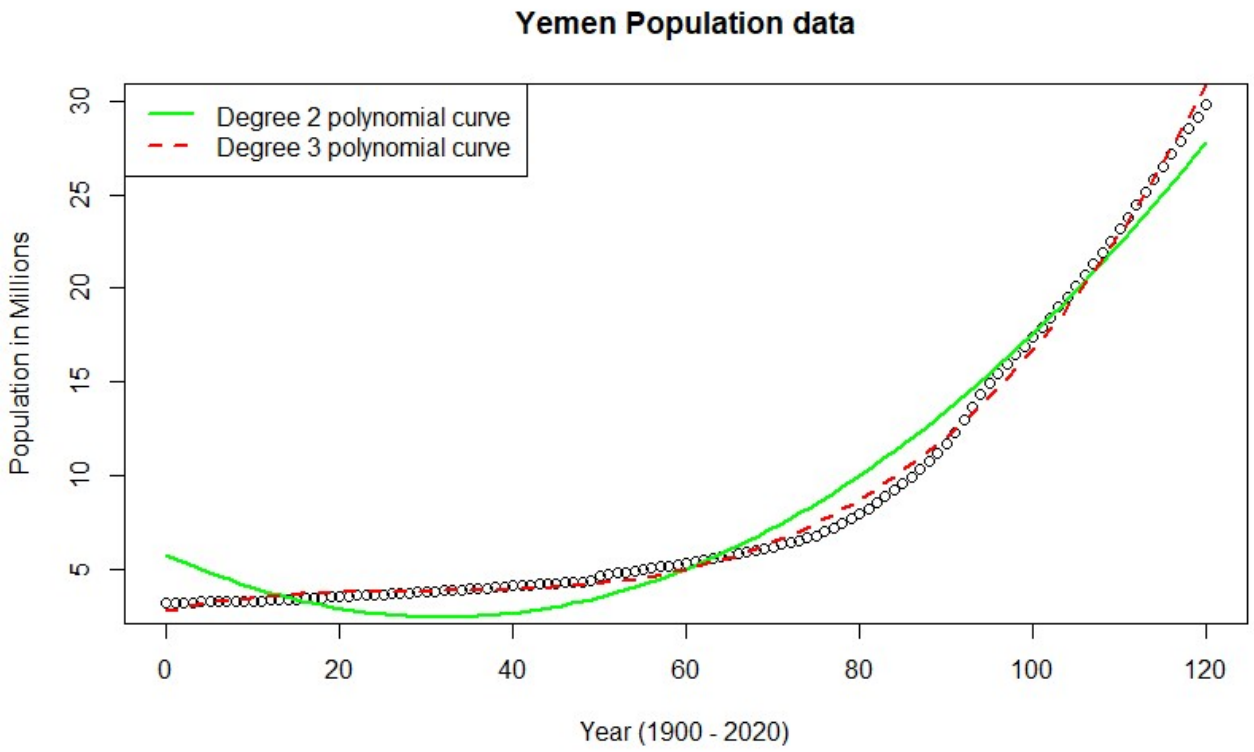
Regression Equation:
Weight = -203.190 + 5.772*Height + -11.615* Gender (Male)
(Input 1 in Gender for Male & 0 for Female)
Coefficient of Determination = (179804.2055-148509.23)/179804.2055 = **17.4%**

Q2 Hanes Data plot with interaction term (gender: height)



Regression Equation:
Weight = -143.998 + 4.854 * Height -145.143* Gender (Male) + 1.998 * Height * Gender
(Input 1 in Gender for Male & 0 for Female)
Coefficient of Determination =(179804.2055-147695.634)/179804.2055 = **17.85%**

Q3. Yemen Data Plot



Degree 2 Polynomial:
Population = 5.763782 -0.209238 * year + 0.003274 * year^2
Coefficient of Determination = **97.13%**

Degree 3 Polynomial:
Population = 2.759e+00 + 9.764e-02 * year - 3.146e-03 * year^2 + 3.567e-05 * year^3
Coefficient of Determination = **99.70%**

R Script

Q1:

```
#Defining Functions
avg <- function(x){
  sum(x)/length(x)
}

#Q1 Hanes Data with Gender Indicators
han <- na.omit(hanes)
plot(x=han$height, y=han$weight,
     xlab="Height (Inches)",
     ylab="Weight (Pounds)",
     main=" Hanes Data with Gender Indicators",
     pch= ifelse(han$gender =="F",20,8),
     col =ifelse(han$gender == "F","red","black"))

#linear regression
lm(han$weight ~ han$height + han$gender)
#regression line for female
abline(a=-203.190, b=5.772, col="red", lwd=2 )
#regression line for male
abline(a=-214.805, b=5.772, col="black", lwd=2 )

#Legend regression lines
legend("topleft", c("Female", "Male"), lwd =2,
       col=c("red","black"))

#calculating Sum of squares & Doefficient of Determination
TotalSumSqHan <- sum((han$weight - avg(han$weight))^2)
ResidualSumSqHan <- sum(lm(han$weight ~ han$height + han$gender)$residuals^2)
CoDHan <- ((TotalSumSqHan - ResidualSumSqHan)/TotalSumSqHan)*100
CoDHan
```

Q2:

```
#Defining Functions
avg <- function(x){
  sum(x)/length(x)
}
#Q2 Hanes Data with gender indicators and interaction term
han <- na.omit(hanes)
plot(x=han$height, y=han$weight,
     xlab="Height (Inches)",
     ylab="Weight (Pounds)",
     main="Hanes Data with gender indicators and interaction terms",
     pch= ifelse(han$gender =="F",20,8),
     col =ifelse(han$gender == "F","red","black"))

#linear regression
lm(han$weight ~ han$height + han$gender + han$gender:han$height)
#regression line for female
abline(a=-143.998, b=4.854, col="red", lwd=2 )
#regression line for male
abline(a=-289.141, b=6.852, col="black", lwd=2 )

#Legend regression lines
legend("topleft", c("Female", "Male"), lwd =2,
       col=c("red","black"))

#calculating Sum of squares & Doefficient of Determination
TotalSumSqHan <- sum((han$weight - avg(han$weight))^2)
ResidualSumSqHan <- sum(lm(han$weight ~ han$height + han$gender +
han$gender:han$height)$residuals^2)
CoDHan <- ((TotalSumSqHan - ResidualSumSqHan)/TotalSumSqHan)*100
CoDHan
```

Q3:

```
# Function Definitions
avg <- function(x){
  sum(x)/length(x)
}
#Loading CSV File for Yemen population from Year 1900 - 2020, year count starting from 0.
Yemen <- read.csv("Yemen.csv")

#plotting scatterplot
plot(x=Yemen$year, y=Yemen$population,
     xlab="Year (1900 - 2020)",
     ylab="Population in Millions",
     main="Yemen Population data",
     pch=1)

#Degree 2 Polynomial
year_2 <- Yemen$year^2
#Degree 3 Polynomial
year_3 <- Yemen$year^3

#Linear regression for degree2 polynomial
lm(Yemen$population ~ Yemen$year + year_2)
deg_2 <- lm(Yemen$population ~ Yemen$year + year_2)
deg_2$coefficients
curve(5.763782  - 0.209238 *x + 0.003274 *x^2 , add=TRUE, col='green',lwd=2, lty =1)

#Linear regression for degree3 polynomial
lm(Yemen$population ~ Yemen$year + year_2 + year_3)
deg_3 <- lm(Yemen$population ~ Yemen$year + year_2 + year_3)
deg_3$coefficients
curve(2.759e+00 + 9.764e-02 *x -3.146e-03 *x^2  +3.567e-05*x^3, add=TRUE,col='red',lwd=2, lty=2
)

#Legends for curve
legend("topleft", c("Degree 2 polynomial curve", "Degree 3 polynomial curve"), lwd =2
,lty=c(1,2),col
       =c("green","red"))

#Sum of squares & Coefficient of Determination Calucations
TotalSumSq <- sum((Yemen$population - avg(Yemen$population))^2)
ResidualSumSq_2 <- sum(lm(Yemen$population ~ Yemen$year + year_2)$residuals^2)
ResidualSumSq_3 <- sum(lm(Yemen$population ~ Yemen$year + year_2 + year_3)$residuals^2)
CoeffD_2 <- (TotalSumSq - ResidualSumSq_2)/TotalSumSq
CoeffD_3 <- (TotalSumSq - ResidualSumSq_3)/TotalSumSq
CoeffD_2
CoeffD_3
```