Assignment 4

Evan Hodges

2025-03-27

Question 1

Part A

```
# Create the vectors
Location <- c("A", "A", "A", "B", "B", "B", "C")
Height <- c(100, 200, 300, 450, 600, 800, 1000)
Distance <- c(253, 337, 395, 451, 495, 534, 573)

# Create the data frame
Galileo <- data.frame(Location, Height, Distance)

# Display the data frame
Galileo
```

```
##   Location Height Distance
## 1        A    100      253
## 2        A    200      337
## 3        A    300      395
## 4        B    450      451
## 5        B    600      495
## 6        B    800      534
## 7        C   1000      573
```

Part B

```
# Compute sample mean, median, variance, and IQR for Distance
mean_distance <- mean(Galileo$Distance)
median_distance <- median(Galileo$Distance)
variance_distance <- var(Galileo$Distance)
iqr_distance <- IQR(Galileo$Distance)

# Display the results
mean_distance
```

```
## [1] 434
```

```
median_distance
```

```
## [1] 451
```

```r
variance_distance
```

```
## [1] 12837
```

```r
iqr_distance
```

```
## [1] 148.5
```

Part C

```r
# Create estimated distance D.Hat and add it to the data frame
Galileo$D.Hat <- 200 + 0.708 * Galileo$Height - 0.000344 * (Galileo$Height^2)

# Create the LO variable: TRUE if estimated distance is lower than measured distance
Galileo$LO <- Galileo$D.Hat < Galileo$Distance

# Display the updated data frame
Galileo
```

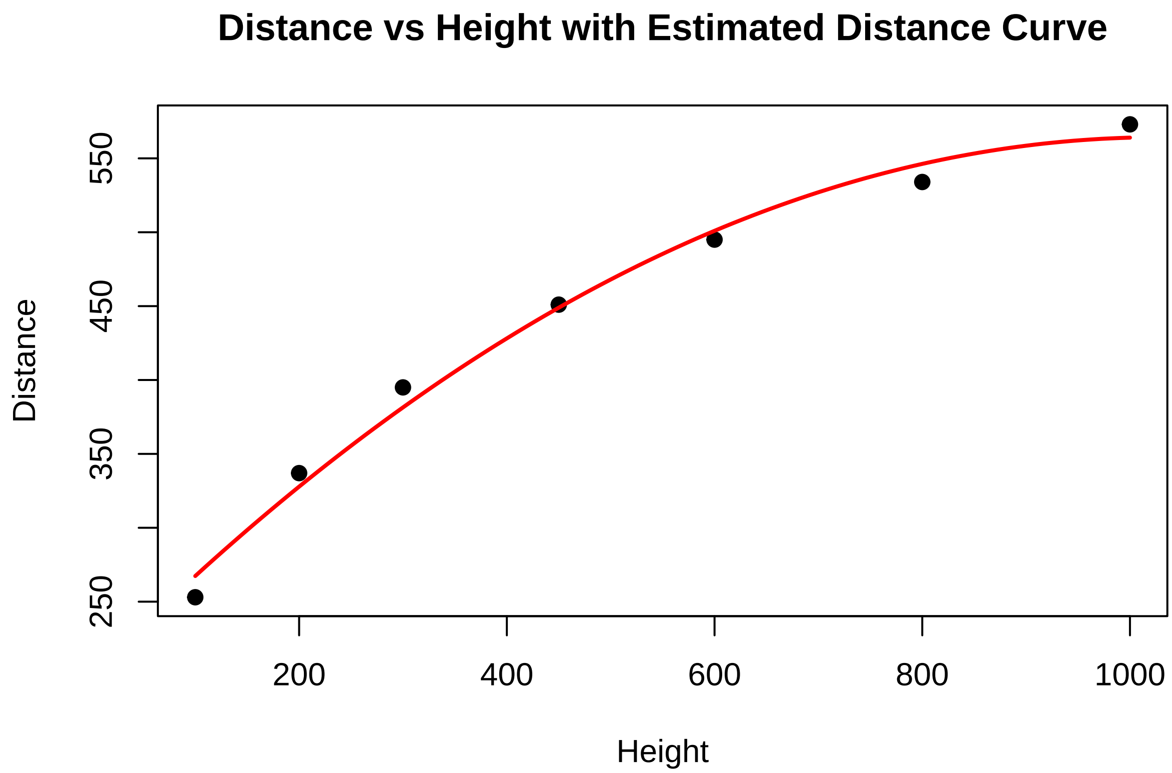
```
##   Location Height Distance  D.Hat    LO
## 1        A    100      253 267.36 FALSE
## 2        A    200      337 327.84  TRUE
## 3        A    300      395 381.44  TRUE
## 4        B    450      451 448.94  TRUE
## 5        B    600      495 500.96 FALSE
## 6        B    800      534 546.24 FALSE
## 7        C   1000      573 564.00  TRUE
```

```r
# Extract the subset where LO is FALSE
Galileo_subset <- Galileo[!Galileo$LO, ]
Galileo_subset
```

```
##   Location Height Distance  D.Hat    LO
## 1        A    100      253 267.36 FALSE
## 5        B    600      495 500.96 FALSE
## 6        B    800      534 546.24 FALSE
```

Part D

```r
# Scatterplot of Distance vs Height
plot(Galileo$Height, Galileo$Distance,
     main = "Distance vs Height with Estimated Distance Curve",
     xlab = "Height", ylab = "Distance", pch = 19)

# Create a sequence of Height values for a smooth curve
height_seq <- seq(min(Galileo$Height), max(Galileo$Height), length.out = 100)
dhat_curve <- 200 + 0.708 * height_seq - 0.000344 * (height_seq^2)

# Overlay the curve for estimated distance
lines(height_seq, dhat_curve, col = "red", lwd = 2)
```



Question 2

Part A

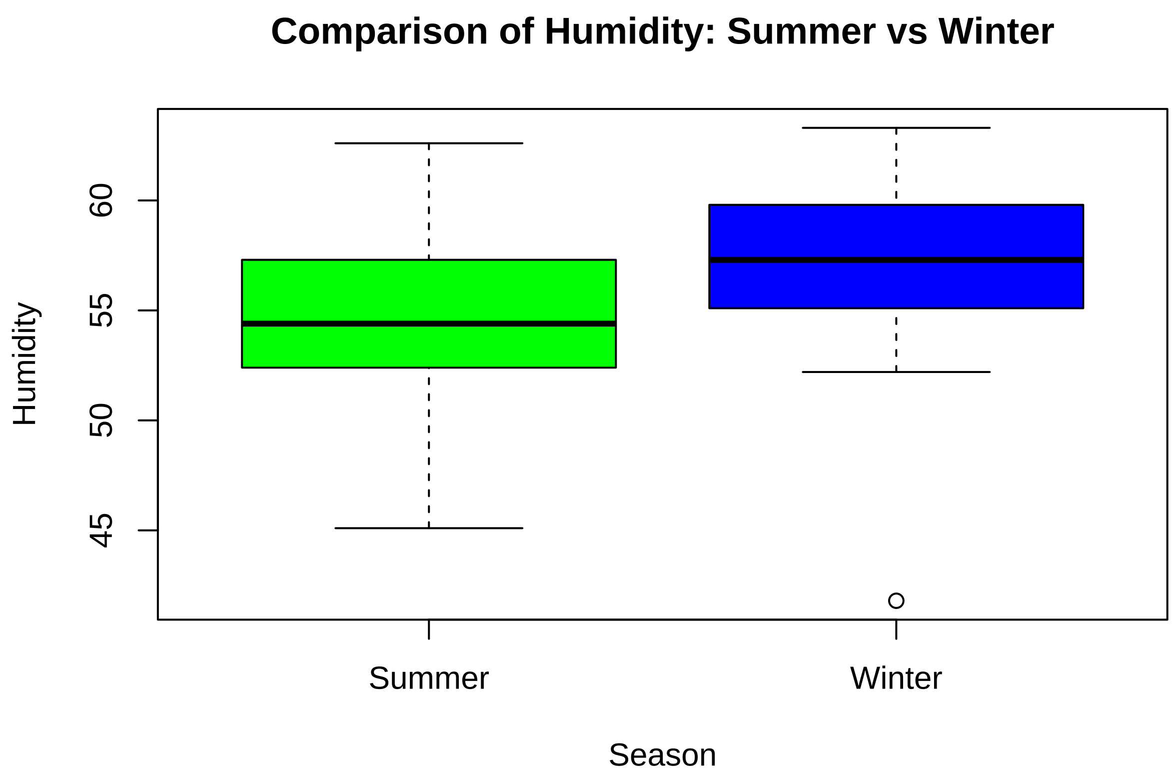
```r
# Read in the data from the CSV file
humidity_data <- read.csv("hw4q2.csv", header = TRUE)

# Inspect the first few rows and column names to verify the data
head(humidity_data)
```

```
##   Humidity Season
## 1     57.2 Summer
## 2     58.1 Summer
## 3     56.5 Summer
## 4     58.6 Summer
## 5     57.4 Summer
## 6     62.6 Summer
```

```r
names(humidity_data)
```

```
## [1] "Humidity" "Season"
```

```r
# Create a boxplot comparing Humidity for Summer and Winter
boxplot(Humidity ~ Season, data = humidity_data,
        main = "Comparison of Humidity: Summer vs Winter",
        xlab = "Season", ylab = "Humidity",
        col = c("green", "blue"))
```

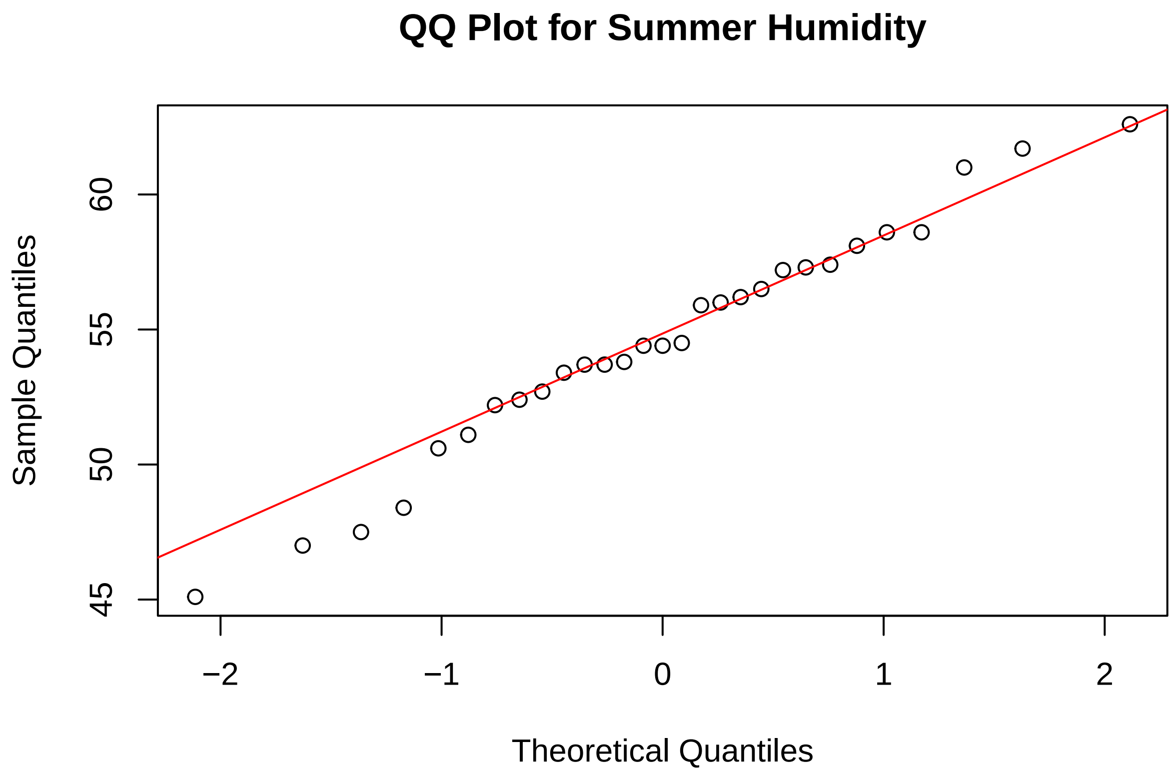


Comments:

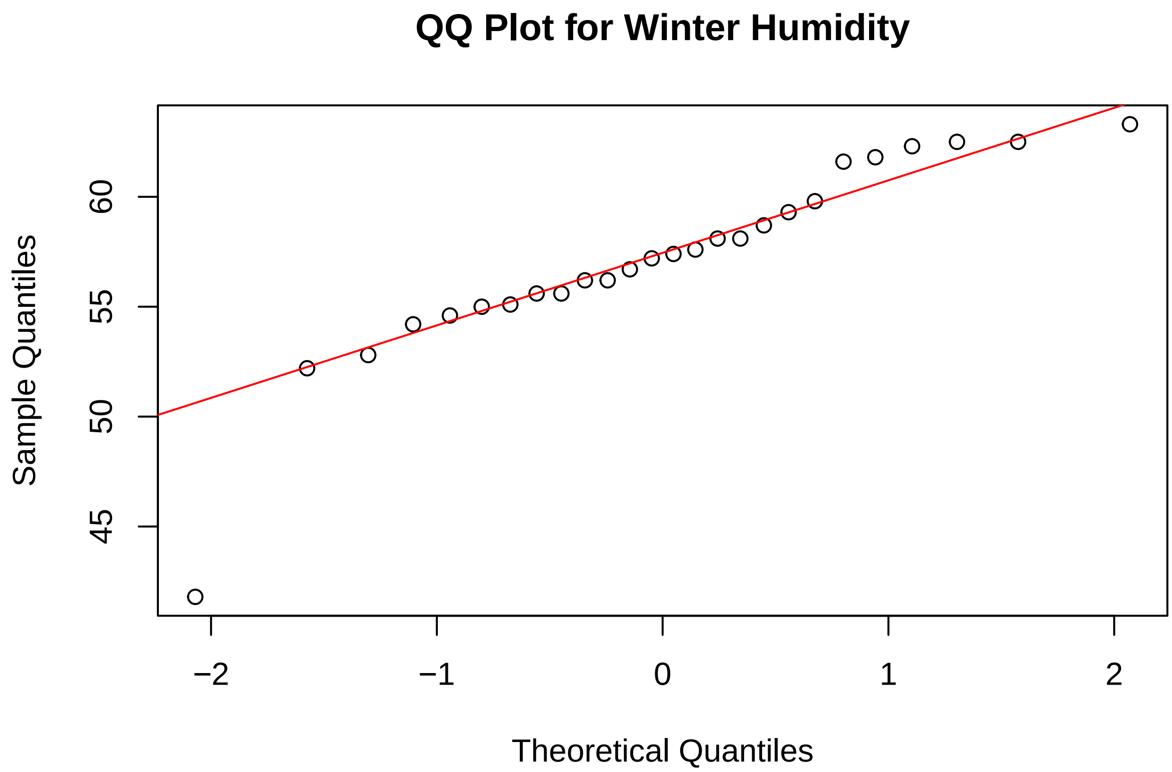
Winter's median is slightly higher than Summer's. Summer's humidity ranges more widely, while Winter's main cluster is tighter but includes a lower outlier. Winter shows slightly higher average humidity; Summer exhibits greater variability.

Part B

```r
# Subset the data by Season
summer_humidity <- subset(humidity_data, Season == "Summer")$Humidity
winter_humidity <- subset(humidity_data, Season == "Winter")$Humidity

# QQ plot for Summer humidity
qqnorm(summer_humidity, main = "QQ Plot for Summer Humidity")
qqline(summer_humidity, col = "red")
```




```
# QQ plot for Winter humidity
qqnorm(winter_humidity, main = "QQ Plot for Winter Humidity")
qqline(winter_humidity, col = "red")
```



Comments:

For the summer plot, the points generally align well with the reference line, suggesting that summer humidity is reasonably normally distributed. Any minor deviations are near the lower tail, indicating a few lower humidity readings that are slightly outside the main distribution. For the winter plot, the points also follow the line for much of the range, but the upper tail shows a slight upward deviation. This suggests a possible right skew or heavier upper tail in the winter humidity distribution compared to summer.

Part C

```r
# Calculate variance and IQR for Summer humidity
var_summer <- var(summer_humidity)
iqr_summer <- IQR(summer_humidity)

# Calculate variance and IQR for Winter humidity
var_winter <- var(winter_humidity)
iqr_winter <- IQR(winter_humidity)

# Print the results
var_summer
```

```
## [1] 18.37044
```

```r
iqr_summer
```

```
## [1] 4.9
```

```
var_winter
```

```
## [1] 19.56086
```

```
iqr_winter
```

```
## [1] 4.45
```

Comments:

Winter humidity has a slightly higher variance than Summer, indicating marginally greater spread in Winter if we measure variability by variance. Summer humidity has a slightly higher IQR than Winter, suggesting marginally greater spread in Summer if we measure variability by the middle 50% of data.
