Aidan Kiser
20 March 2024
STAT 3010
Xia

Assignment 4

All screenshots that are included are from my own R Script file. You can access it with this link: https://github.com/kiseraidan/STAT-3010/blob/main/Assignment%204/Assignment4.R

1. (a).

Code:

```
15  Galileo <- data.frame(Location, Height, Distance)
16  print(Galileo)
```

Output:

	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

(b).

Code:

```r
mean_distance <- mean(Galileo$Distance)
print(paste("Mean:", mean_distance))

# Median
median_distance <- median(Galileo$Distance)
print(paste("Median:", median_distance))

# Variance
variance_distance <- var(Galileo$Distance)
print(paste("Variance:", variance_distance))

# IQR
iqr_distance <- IQR(Galileo$Distance)
print(paste("IQR:", iqr_distance))
```

Output:

```
[1] "Mean: 434"
[1] "Median: 451"
[1] "Variance: 12837"
[1] "IQR: 148.5"
>
```

(c).

Code:

```r
# Calculate D.Hat
Galileo$D.Hat <- 200 + 0.708 * Galileo$Height - 0.000344 * Galileo$Height^2

# Create LO variable
Galileo$LO <- Galileo$D.Hat < Galileo$Distance

# Extract a subset removing observations where estimated distance is lower than the measured distance
Galileo_subset <- Galileo[!Galileo$LO, ]

# Optionally, remove the D.Hat and LO columns for display purposes, if you wish
Galileo_subset <- Galileo_subset[, !(names(Galileo_subset) %in% c("D.Hat", "LO"))]

# Show the contents of this filtered dataframe
print(Galileo_subset)

```

Output:

	Location	Height	Distance
1	A	100	253
5	B	600	495
6	B	800	534

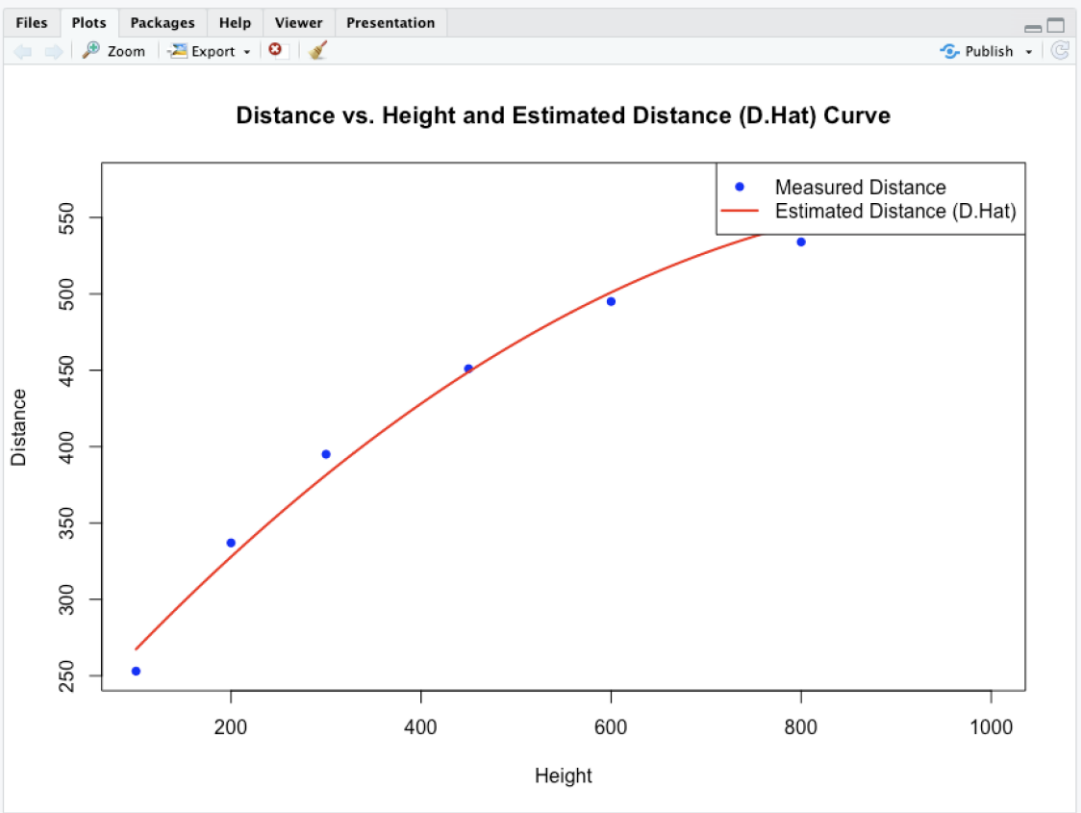
```
>
```

(d).

Code:

```r
# Plotting Distance versus Height
plot(Galileo$Height, Galileo$Distance,
     xlab = "Height", ylab = "Distance",
     main = "Distance vs. Height and Estimated Distance (D.Hat) Curve",
     col = "blue", pch = 16)

# Generating points for the D.Hat curve
height_range <- seq(min(Galileo$Height), max(Galileo$Height), length.out = 500)
d_hat_curve <- 200 + 0.708 * height_range - 0.000344 * height_range^2

# Overlaying the curve of D.Hat
lines(height_range, d_hat_curve, col = "red", lwd = 2)

# Adding a legend
legend("topright", legend = c("Measured Distance", "Estimated Distance (D.Hat)"),
       col = c("blue", "red"), pch = c(16, NA), lty = c(NA, 1), lwd = c(NA, 2))
```

Output:



2. (a).

Code:

```
data <- read.csv("/Users/aidankiser/Documents/Spring 2024 Classes/Statistics for Engineers & Scientists/Assignments/Assignment 4/hw4q2.c

# Generating the comparison boxplot for Summer and Winter humidity
boxplot(Humidity ~ Season, data = data,
        xlab = "Season", ylab = "Humidity (%)",
        main = "Comparison of Humidity in Summer and Winter",
        col = c("lightblue", "lightgreen"))

```

Output:

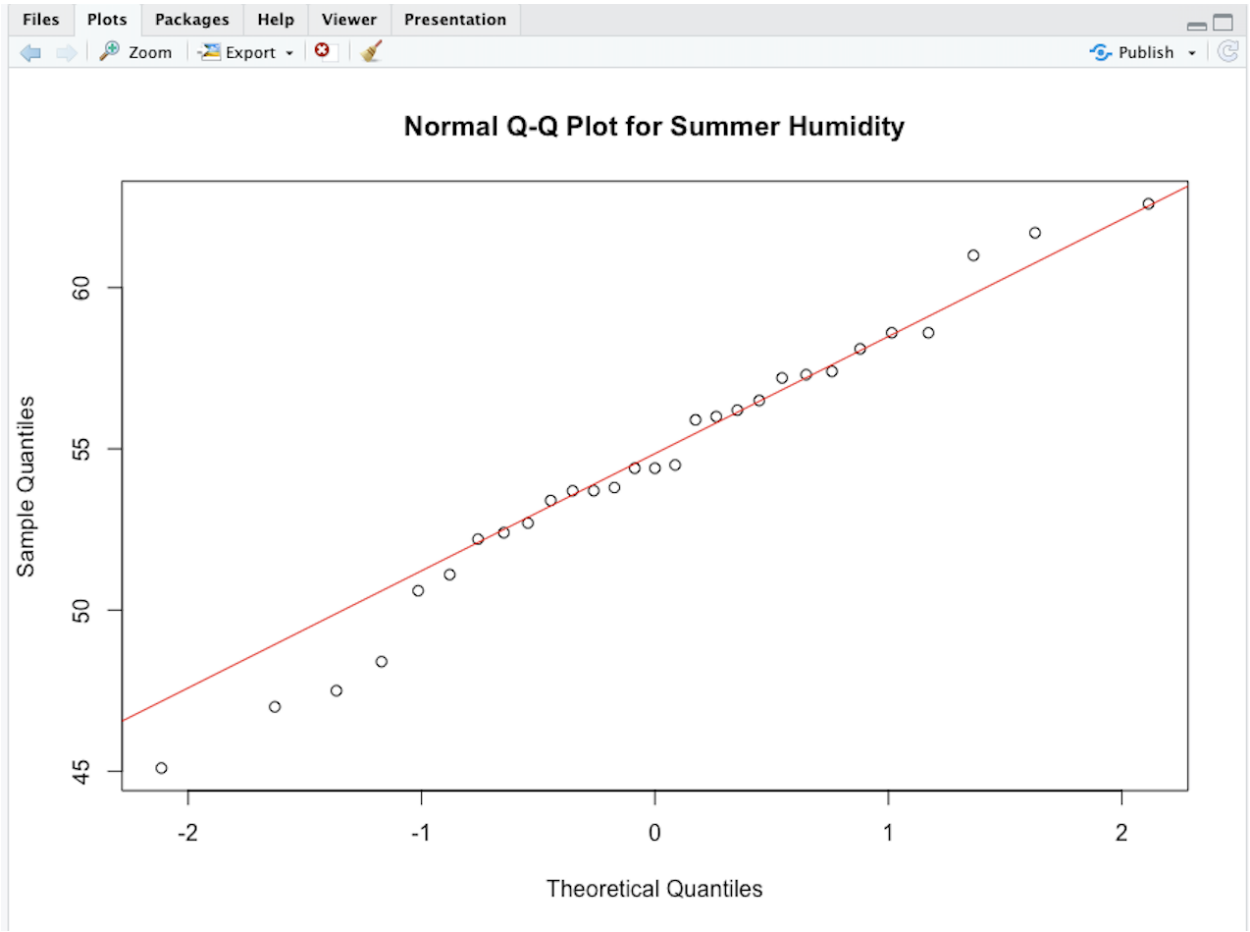
Observations:

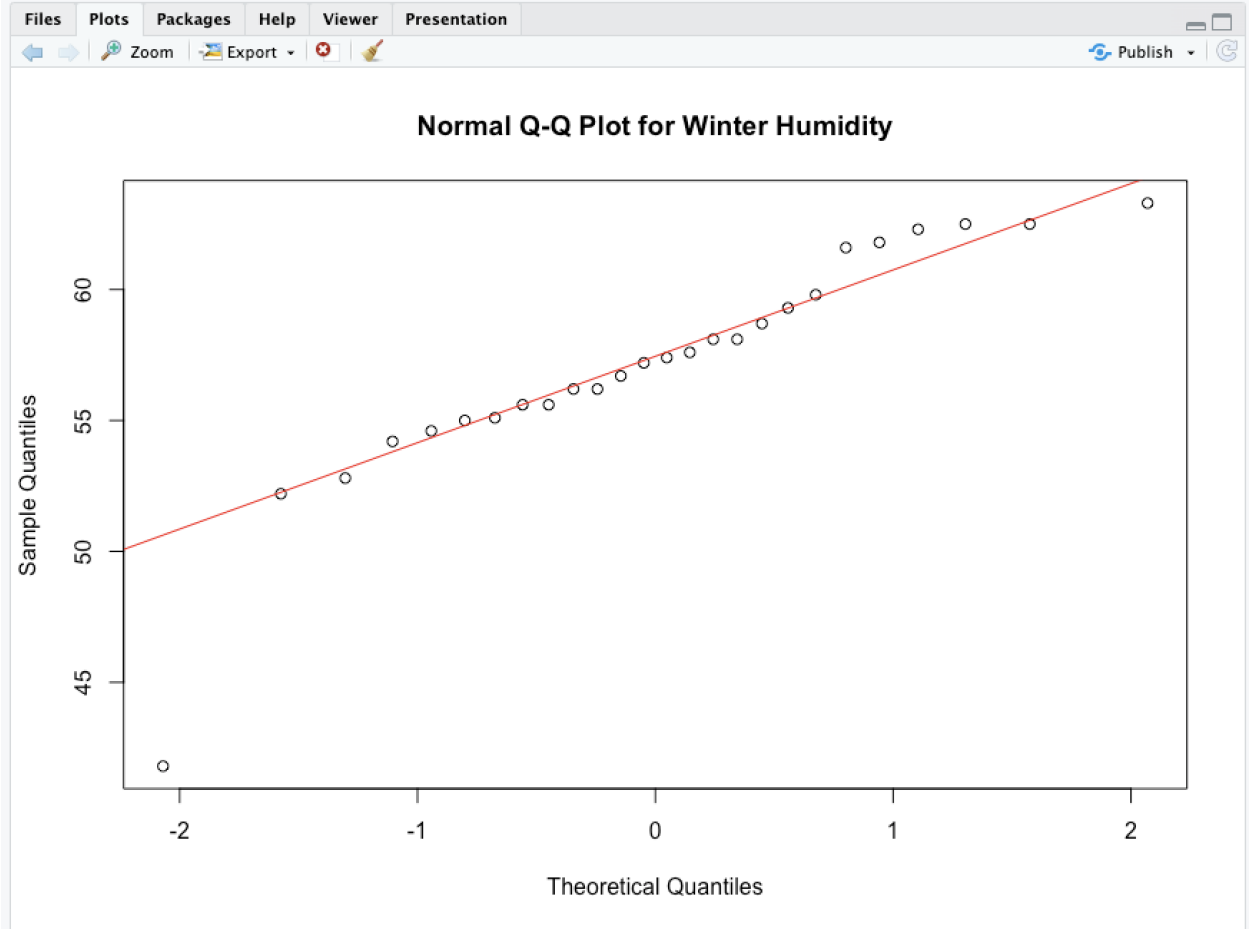
- The median humidity in Summer appears to be higher than in Winter, indicating generally more humid conditions during the Summer months.
- The spread of humidity values (as indicated by the interquartile range, IQR) is broader in Summer than in Winter, suggesting greater variability in humidity levels during Summer.
- Both seasons show outliers, but Winter seems to have outliers on the lower end, indicating occasional very dry conditions.
- The overall range (from minimum to maximum) of humidity levels is wider in Summer, reflecting both the higher variability and the extremes of humidity experienced during this season.

(b).
Code:

```r
# Load necessary library
library(ggplot2)

# Data filtering for each season
summer_humidity <- data[data$Season == 'Summer', 'Humidity']
winter_humidity <- data[data$Season == 'Winter', 'Humidity']

# Plot for Summer Humidity
qqnorm(summer_humidity, main = "Normal Q-Q Plot for Summer Humidity")
qqline(summer_humidity, col = "red")

# Plot for Winter Humidity
qqnorm(winter_humidity, main = "Normal Q-Q Plot for Winter Humidity")
qqline(winter_humidity, col = "red")
```

Output:





Observations:

- Both distributions show some deviation from normality, implying that the relative humidity data for both seasons may not perfectly fit a normal distribution. This could be due to the presence of outliers, skewness in the data, or other factors that affect the distribution of humidity.
- The deviations are more noticeable for Summer humidity, which could indicate more extreme values or greater variability in humidity levels during the Summer months.
- The information from these plots is crucial for statistical analyses that assume normality. If data significantly deviate from normality, alternative methods or transformations might be necessary to properly analyze the data.

(c).
Code:

```r
# Calculating variance for each season
variance_summer <- var(summer_humidity)
variance_winter <- var(winter_humidity)

# Calculating IQR for each season
iqr_summer <- IQR(summer_humidity)
iqr_winter <- IQR(winter_humidity)

# Printing the results
cat("Variance - Summer Humidity:", variance_summer, "\n")
cat("Variance - Winter Humidity:", variance_winter, "\n")
cat("IQR - Summer Humidity:", iqr_summer, "\n")
cat("IQR - Winter Humidity:", iqr_winter, "\n")
```

Output:

```
Variance - Summer Humidity: 18.37044
Variance - Winter Humidity: 19.56086
IQR - Summer Humidity: 4.9
IQR - Winter Humidity: 4.45
>
```

Observations:

- The variances for Summer and Winter humidity are relatively close, with Winter having a slightly higher variance than Summer. This suggests that the spread of humidity values around the mean is somewhat greater in Winter than in Summer.
- The IQRs are also quite similar, with Summer having a marginally higher IQR than Winter. The IQR reflects the spread of the middle 50% of the data, indicating that the core range of humidity levels is slightly broader in Summer.
- These metrics imply that while there are differences in the spread and central dispersion of humidity values between the seasons, they are not drastically different. Both seasons show a moderate level of variability in humidity levels.

- The higher variance in Winter might be influenced by the occasional very dry conditions (as seen in the boxplot outliers), whereas the slightly broader IQR in Summer suggests a wider spread of humidity conditions within the most common range of values.