Lab 1

Christopher Hainzl

2023-02-04

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Question 1

```
noises <- c(55.3, 55.3, 55.3, 55.9, 55.9, 55.9, 55.9, 56.1, 56.1, 56.1, 56.1,
56.1, 56.1, 56.8, 56.8, 57.0, 57.0, 57.0, 57.8, 57.8, 57.8, 57.9,
57.9, 57.9, 58.8, 58.8, 58.8, 59.8, 59.8, 59.8, 62.2, 62.2, 63.8,
63.8, 63.8, 63.9, 63.9, 63.9, 64.7, 64.7, 64.7, 65.1, 65.1, 65.1,
65.3, 65.3, 65.3, 65.3, 67.4, 67.4, 67.4, 67.4, 68.7, 68.7, 68.7,
68.7, 69.0, 70.4, 70.4, 71.2, 71.2, 71.2, 73.0, 73.0, 73.1, 73.1,
74.6, 74.6, 74.6, 74.6, 79.3, 79.3, 79.3, 79.3, 83.0, 83.0, 83.0)
```

```
# Compute sample mean, median, standard deviation and inter-quartile range for noise data.
mean(noises)
```

```
## [1] 64.88701
```

```
median(noises)
```

```
## [1] 64.7
```

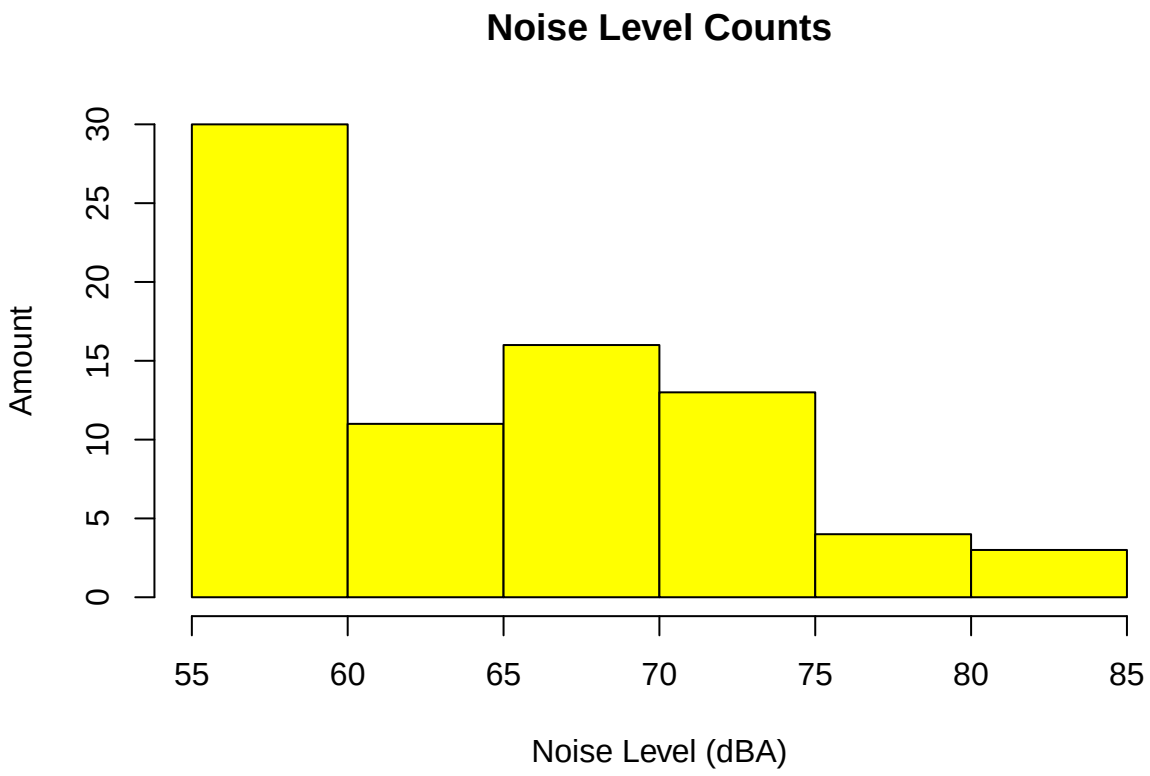
```
sd(noises)
```

```
## [1] 7.802671
```

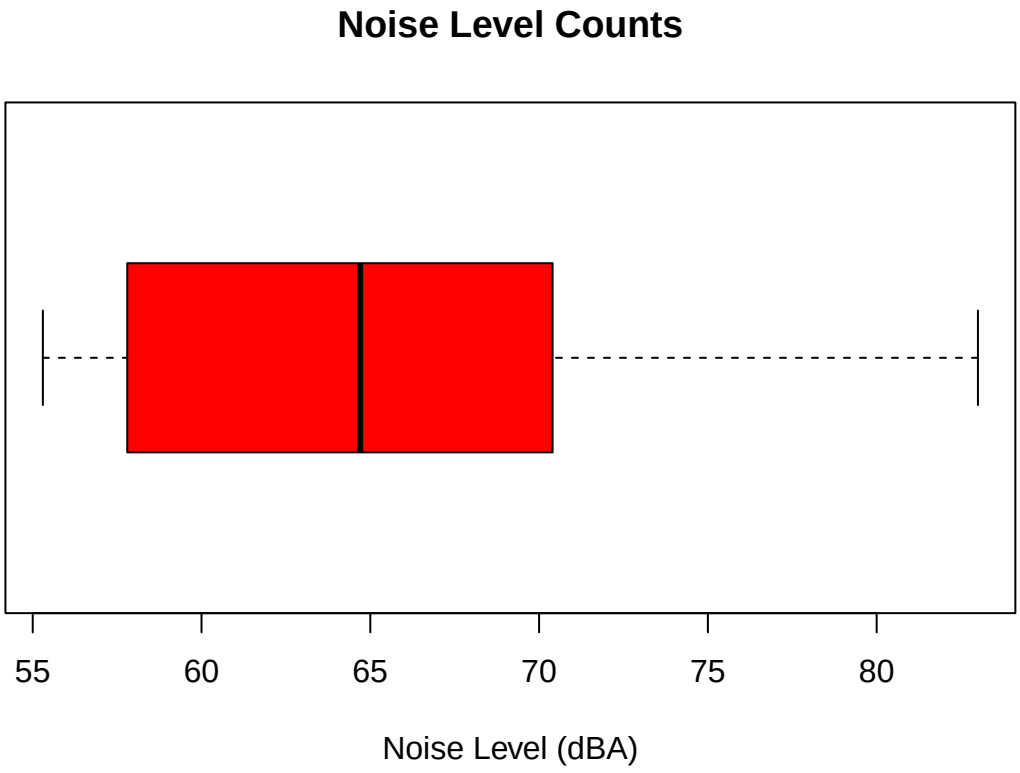
```
IQR(noises)
```

```
## [1] 12.6
```

```
# Generate histogram and boxplot
hist(noises, xlab = "Noise Level (dBA)", ylab = "Amount", col = "yellow", main = "Noise Level Counts")
```




```
boxplot(noises, xlab = "Noise Level (dBA)", col = "red", main = "Noise Level Counts",
horizontal = TRUE)
```



The mean and median of the data are approximately 64.89 and 64.7, respectively. The standard deviation and inter-quartile range of the data are approximately 7.80 and 12.6, respectively. The histogram appears to be bimodal and indicates that the data is right-skewed. The boxplot suggests that there are no outliers.

Question 2

Question 2a

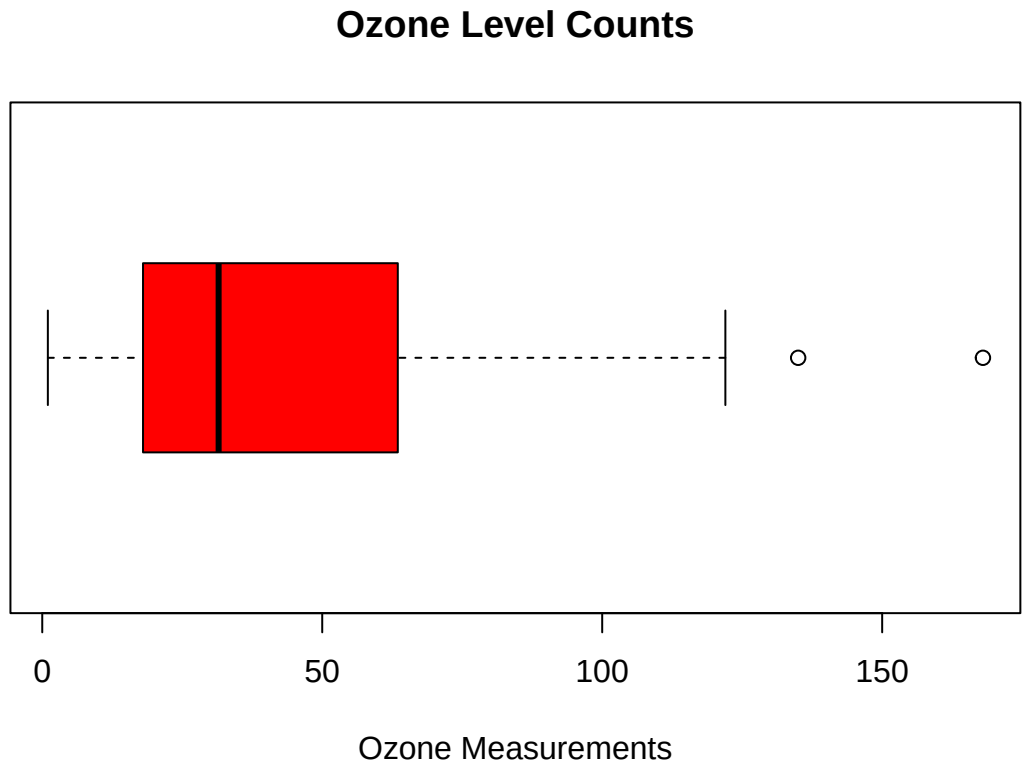
```
# Compute sample standard deviation
sd(airquality$Ozone, na.rm = TRUE)
```

```
## [1] 32.98788
```

```
# Compute sample mean
mean(airquality$Ozone, na.rm = TRUE)
```

```
## [1] 42.12931
```

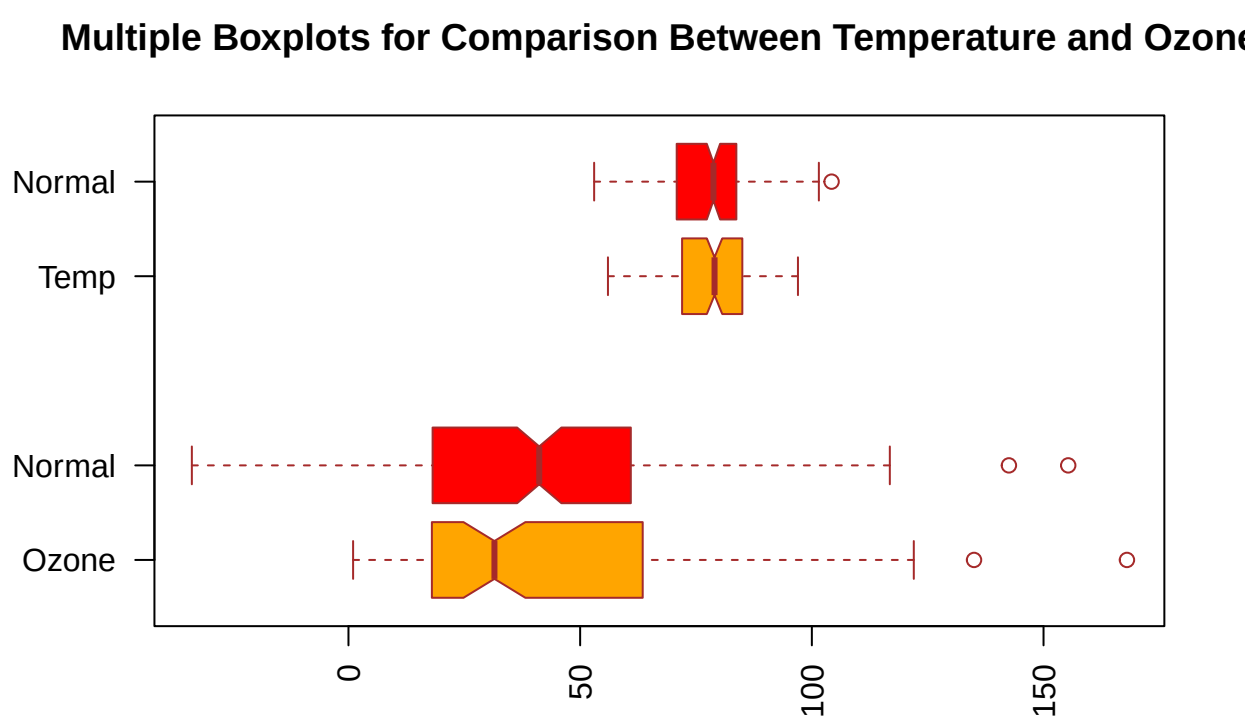
```
# Generate horizontal boxplot
boxplot(airquality$Ozone, xlab = "Ozone Measurements", col = "red", main = "Ozone Level Counts",
        horizontal = TRUE)
```



The standard deviation of the ozone data is approximately 33.00, while the sample mean is approximately 42.13. The boxplot shows that there are 2 outliers in the ozone data.

Question 2b

```
ozone_norm <- rnorm(200, mean(airquality$Ozone, na.rm = TRUE), sd(airquality$Ozone, na.rm = TRUE))
temp_norm <- rnorm(200, mean(airquality$Temp, na.rm = TRUE), sd(airquality$Temp, na.rm = TRUE))
boxplot(airquality$Ozone, ozone_norm, airquality$Temp, temp_norm,
main = "Multiple Boxplots for Comparison Between Temperature and Ozone",
at = c(1,2,4,5),
names = c("Ozone", "Normal", "Temp", "Normal"),
las = 2,
col = c("orange","red"),
border = "brown",
horizontal = TRUE,
notch = TRUE)
```

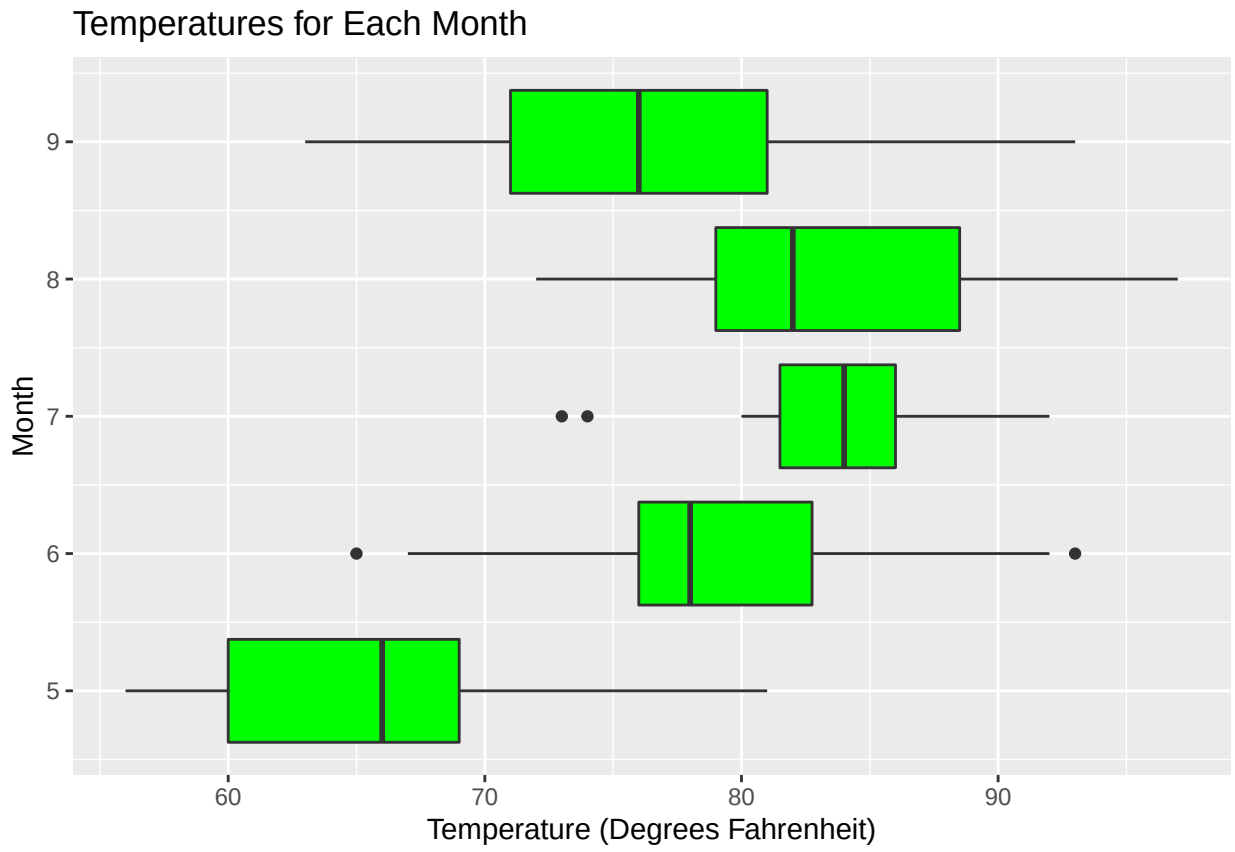


Question 2c

Only the temperature data appears to have been approximately drawn from a normal distribution.

Question 2d

```
# Generate a boxplot for each month in the data
ggplot(data = airquality, aes(x = Month, y = Temp, group = Month)) +
  geom_boxplot(fill = "green") +
       labs(
         title = "Temperatures for Each Month",
         x = "Month",
         y = "Temperature (Degrees Fahrenheit)"
       ) +
  coord_flip()
```




```
june <- airquality %>% filter(airquality$Month == 6)
summary(june$Temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.00   76.00   78.00   79.10   82.75   93.00
```

```
IQR(june$Temp)
```

```
## [1] 6.75
```

```
july <- airquality %>% filter(airquality$Month == 7)
summary(july$Temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    73.0    81.5    84.0    83.9    86.0    92.0
```

```
IQR(july$Temp)
```

```
## [1] 4.5
```

```
boxplot.stats(july$Temp)$out
```

```
## [1] 73 74
```

Question 2e

Using the boxplots, I observed the following:

- The month of September appeared to have the most normal distribution compared to the others.

- The difference in temperature appears to be most significant between the months of May and July.
- There was more variance in temperature when considering every month except July.
- There are outliers present only in the data for the months of June and July. June has two outliers, which are 65 and 93. July also has two outliers, which are 73 and 74.