lab2

Ayimen H.

2025-01-24

```r
library(readxl)

clouds <- read_excel("clouds.xlsx")

unseeded_data <- clouds[clouds$Treatment == "Unseeded", "Rainfall"]
unseeded_rainfall <- unseeded_data$Rainfall

unseeded_mean <- mean(unseeded_rainfall)
unseeded_std_dev <- sd(unseeded_rainfall)
unseeded_five_num <- fivenum(unseeded_rainfall)

seeded_data <- clouds[clouds$Treatment == "Seeded", "Rainfall"]
seeded_rainfall <- seeded_data$Rainfall

seeded_mean <- mean(seeded_rainfall)
seeded_std_dev <- sd(seeded_rainfall)
seeded_five_num <- fivenum(seeded_rainfall)

cat("Mean:", unseeded_mean, "\n")
```

```
## Mean: 164.5885
```

```r
cat("Standard Deviation:", unseeded_std_dev, "\n")
```

```
## Standard Deviation: 278.4264
```

```r
cat("Five-Number Summary:", paste(unseeded_five_num, collapse = " "), "\n")
```

```
## Five-Number Summary: 1 24.4 44.2 163 1202.6
```

```r
cat("Mean:", seeded_mean, "\n")
```

```
## Mean: 441.9846
```

```r
cat("Standard Deviation:", seeded_std_dev, "\n")
```

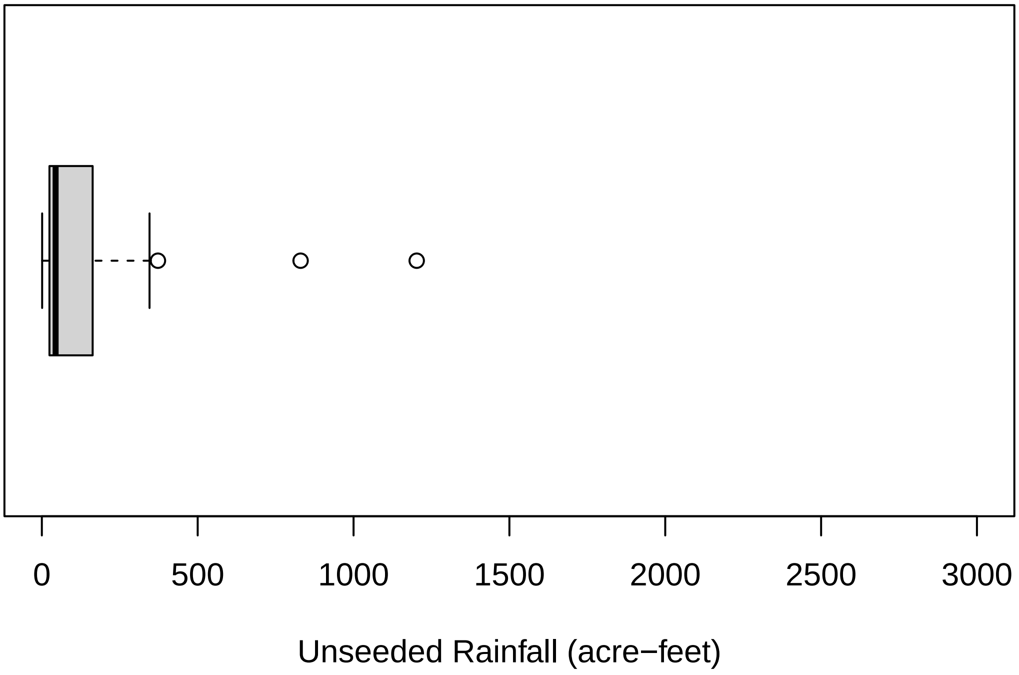
```
## Standard Deviation: 650.7872
```

```r
cat("Five-Number Summary:", paste(seeded_five_num, collapse = " "), "\n")
```

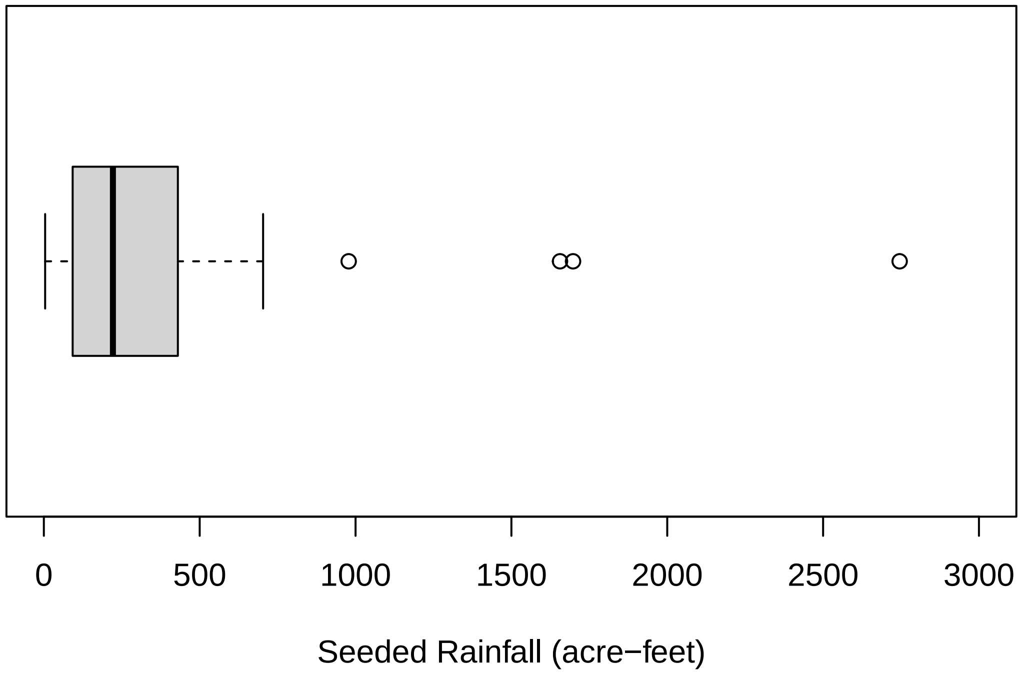
```
## Five-Number Summary: 4.1 92.4 221.6 430 2745.6
```

1.a I would say the five number summary is the most appropriate way of summarizing the data because it gives the most information. The highest and lowest values as well as the median gives us a picture into what is taking place within the data.

```r
library(readxl)

clouds <- read_excel("clouds.xlsx")

unseeded_data <- clouds[clouds$Treatment == "Unseeded", "Rainfall"]
seeded_data <- clouds[clouds$Treatment == "Seeded", "Rainfall"]

unseeded_rainfall <- unseeded_data$Rainfall
seeded_rainfall <- seeded_data$Rainfall

boxplot(unseeded_rainfall, horizontal = TRUE,
        xlab = "Unseeded Rainfall (acre-feet)", ylim = c(0, 3000))
```



```r
boxplot(seeded_rainfall, horizontal = TRUE,
        xlab = "Seeded Rainfall (acre-feet)", ylim = c(0, 3000))
```



1.b As for the effectiveness of cloud seeding, based on the box plots I would say seeding has increased both the average and the extreme acre footage of rain. I set the y limits equivalent to one another to make the comparison easier to see and I could clearly see the black line representing the mean is further right (larger in value) on the seeded box plot than the unseeded one.

```
library(readxl)

clouds <- read_excel("clouds.xlsx")
clouds$log_Rainfall <- log(clouds$Rainfall + 1)

unseeded_log <- clouds$log_Rainfall[clouds$Treatment == "Unseeded"]

unseeded_log_mean <- mean(unseeded_log)
unseeded_log_std_dev <- sd(unseeded_log)
unseeded_log_five_num <- fivenum(unseeded_log)

seeded_log <- clouds$log_Rainfall[clouds$Treatment == "Seeded"]

seeded_log_mean <- mean(seeded_log)
seeded_log_std_dev <- sd(seeded_log)
seeded_log_five_num <- fivenum(seeded_log)

cat("Mean:", unseeded_log_mean, "\n")
```

```
## Mean: 4.051205
```

```
cat("Standard Deviation:", unseeded_log_std_dev, "\n")
```

```
## Standard Deviation: 1.547985
```

```
cat("Five-Number Summary:", paste(unseeded_log_five_num, collapse = " "), "\n")
```

```
## Five-Number Summary: 0.693147180559945 3.23474917402449 3.80873965067343 5.0998664278242 7.0930723447
```

```
cat("Mean:", seeded_log_mean, "\n")
```

```
## Mean: 5.155938
```

```
cat("Standard Deviation:", seeded_log_std_dev, "\n")
```

```
## Standard Deviation: 1.563609
```

```
cat("Five-Number Summary:", paste(seeded_log_five_num, collapse = " "), "\n")
```

```
## Five-Number Summary: 1.62924053973028 4.5368913452348 5.40094919374141 6.06610809010375 7.91811906209
```

2.a I would still say the five number summary is the most appropriate way of summarizing the data because it gives us the most information. The highest and lowest values alongside the median shows us the range (not the mathematical term) of rain fall. It is however less effective because of the normalization (I think that's the term)

```
income_data <- read.csv("top_100_income_zcta.csv", check.names = FALSE)
income_data$PerCapitaIncome <- as.numeric(income_data$`Per Capita Income`)


income_mean <- mean(income_data$PerCapitaIncome, na.rm = TRUE)
income_std_dev <- sd(income_data$PerCapitaIncome, na.rm = TRUE)
income_five_num <- fivenum(income_data$PerCapitaIncome)

cat("Mean:", income_mean, "\n")
```

```
## Mean: 96399.25
```

```
cat("Standard Deviation:", income_std_dev, "\n")
```

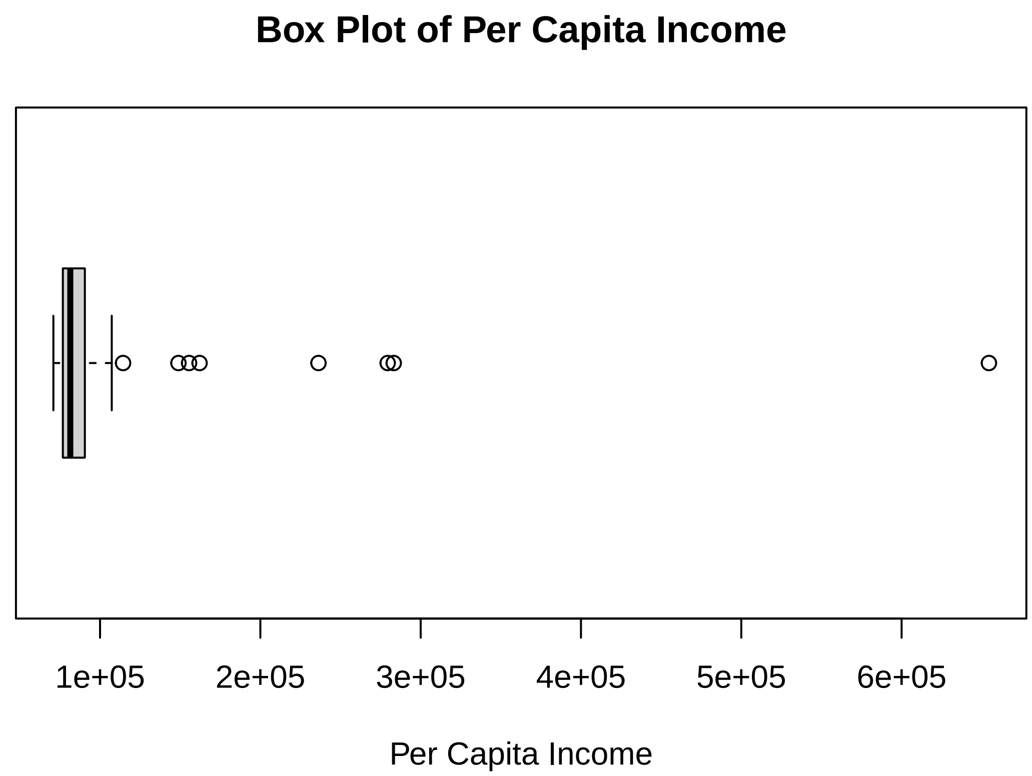
```
## Standard Deviation: 66264.42
```

```
cat("Five-Number Summary:", paste(income_five_num, collapse = " "), "\n")
```

```
## Five-Number Summary: 70878 76787 81433 90526.5 654485
```

3.a An observation I made based off of the summaries of the data is the sheer difference in income especially towards the higher end of the income. Based off looking at five number summary I can tell that the average income is closer to the lower quartile and lowest value. This means there are probably lots of outliers on the right side or richer part of the data.

```
income_data <- read.csv("top_100_income_zcta.csv", check.names = FALSE)
income_data$PerCapitaIncome <- as.numeric(income_data$`Per Capita Income`)

boxplot(income_data$PerCapitaIncome, main = "Box Plot of Per Capita Income",
        xlab = "Per Capita Income", horizontal = TRUE)
```



```
q1 <- quantile(income_data$PerCapitaIncome, 0.25, na.rm = TRUE)
q3 <- quantile(income_data$PerCapitaIncome, 0.75, na.rm = TRUE)

IQR <- q3 - q1

lower_bound <- q1 - 1.5 * IQR
upper_bound <- q3 + 1.5 * IQR

outliers <- income_data$PerCapitaIncome[income_data$PerCapitaIncome < lower_bound
                                        | income_data$PerCapitaIncome > upper_bound]

num_outliers <- length(outliers)

cat("Number of outliers:", num_outliers, "\n")
```

```
## Number of outliers: 8
```

3.b The single most extreme outlier is 654485 which belongs to Montchanin, Delaware with the ZTCA 19710. The challenge with visualizing a data set and having an outlier this extreme is that it squishes the rest of plot to a point where you can not get any meaningful information from it. An idea to address that is maybe a sort of trimmed mean but for the box plot. Meaning a box plot that doesn't contain the extreme outliers. Another idea might be doing log on the values again but I'm not sure how that effects box plots