

Homework 8 Boxplots

Carey Kopeikin and Steph Neul

9/21/2023

Go to the webpage: https://www.openintro.org/data/index.php?data=ssd_speed. **Read the description of the data set before starting!**

If you don't know what a SSD drive is, you can read more about it on Wikipedia: https://en.wikipedia.org/wiki/Solid-state_drive or do your own googling

1. Read in the data found on `ssd_speed.csv` and print out the first six rows.

```
ssd.speed = read.csv("ssd_speed.csv")
head(ssd.speed)
```

```
##      brand      model samples form_factor nvme read write
## 1 Corsair Force MP600   11526         m.2    1  1958  3144
## 2 Samsung    840 Evo   16888         2.5    0   470   389
## 3 Samsung    960 Evo   25990         m.2    1  1798  1562
## 4 Samsung    850 Pro   14690         2.5    0   476   418
## 5 Samsung    970 Pro   21981         m.2    1  2327  2056
## 6 Samsung    860 Pro    2785         2.5    0   481   437
```

2. Use R to find the 5-number summary of average read speeds.

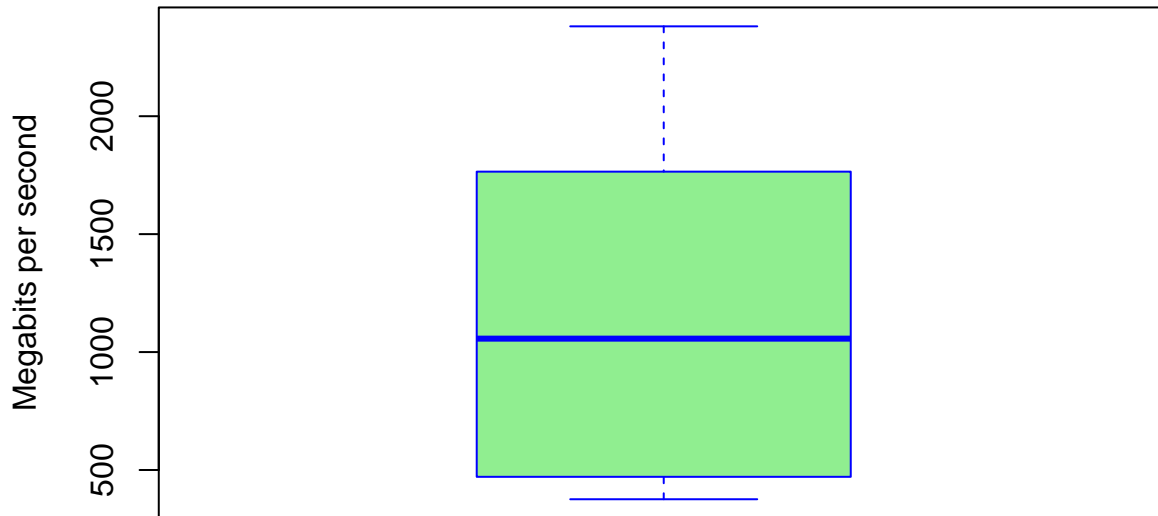
```
summary(ssd.speed$read)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  376.0   471.2  1057.0  1129.0  1764.0  2381.0
```

3. Create a box plot of the average read speeds. Title the plot and label the axis. Make sure that your labels and title make sense. Also, make the colors look nice - I will not give full credit if your graph is black and grey and white. Refer to the R colors png file if you want to see the colors available.

```
boxplot(ssd.speed$read,
        main = "Average Read Speeds of SSDs",
        ylab = "Megabits per second",
        col = "lightgreen",
        border = "blue",
        outcol = "red"
)
```

Average Read Speeds of SSDs



4. Where would the “invisible fences” be on this boxplot? (the dotted lines that we drew on the boxplot that we made by hand - we called them “upper fence” and “lower fence”, and we talked about how we would erase them at the end). Show your work! You should use R to perform any necessary calculations.

```
1764 + 1.5*(1764-471.2)
```

```
## [1] 3703.2
```

```
471.2 - 1.5*(1764-471.2)
```

```
## [1] -1468
```

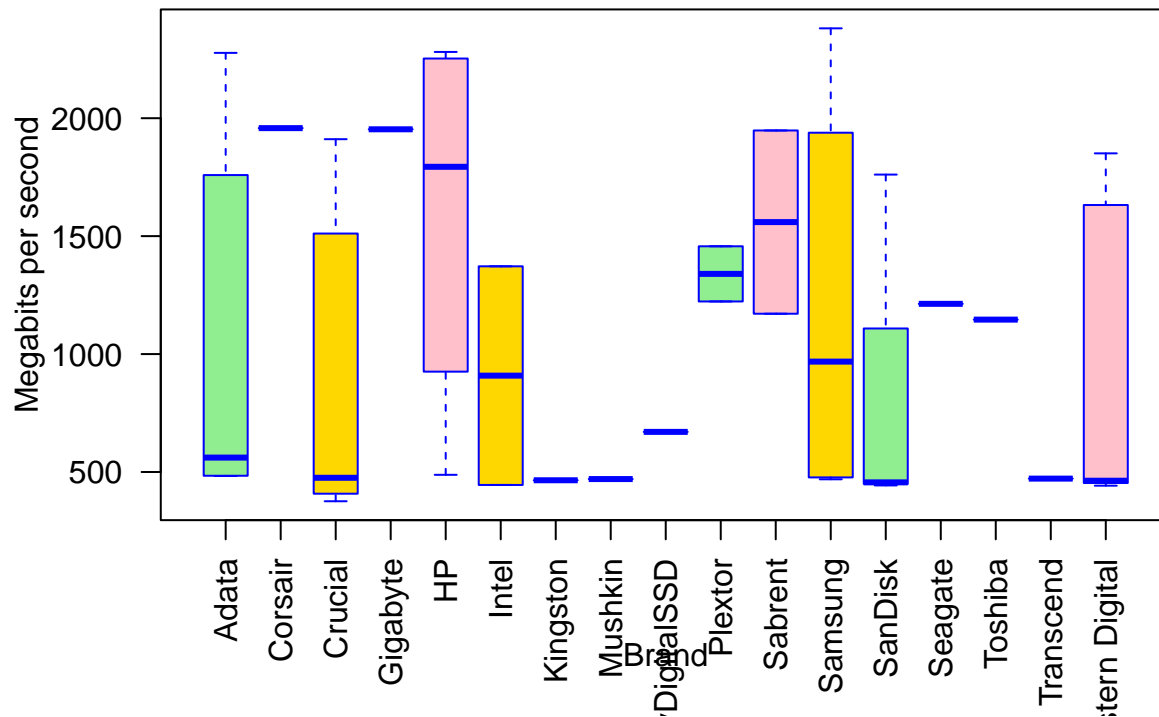
5. Create a boxplot of the **brands** variable. It can be a simple boring boxplot without labels or colors.
6. Why did you get an error when trying to do question 5? Then you should erase your code for problem 5 otherwise this rmd may not properly knit to PDF.

I got an error because the **brands** variable is of type character. R doesn't know how to represent non-numeric variables in a boxplot!

7. Create a boxplot of average **read** speeds by brand (like we did for avocados by type and by region). Include a title, and make sure that all of the axis labels make sense and that every label is visible. Include all titles and labels. Make sure that the colors look nice. *Hint: you can (and should!) turn the axis labels vertical using code found in the Rmd file from class. Look at the graph that we did for avocado average price by region.*

```
boxplot(ssd.speed$read~ssd.speed$brand,
  main = "Average Read Speeds of SSDs",
  xlab = "Brand",
  ylab = "Megabits per second",
  col = c("lightgreen", "pink", "gold"),
  border = "blue",
  outcol = "red",
  las = 2,
)
```

Average Read Speeds of SSDs



8. Write at least 3 good observations about the boxplot you created in problem 7. Each observation should be about a different aspect of the graph. Try to use vocabulary that we talked about in class.

There are no outliers among all of the boxplots. I don't have any other observations; I was a bit challenged by this problem.

9. Of all of the brands represented in this data set, Samsung has greatest number of different models of SSD. By subsetting, create a new dataframe that only contains data for SSDs from the brand Samsung. Then, find the mean read speed for Samsung SSDs.

```
ssd.speed.samsung = ssd.speed[ ssd.speed$brand == "Samsung", ]
head(ssd.speed.samsung)
```

```
##      brand  model samples form_factor nvme read write
## 2 Samsung 840 Evo  16888         2.5   0  470   389
## 3 Samsung 960 Evo  25990         m.2   1 1798  1562
## 4 Samsung 850 Pro  14690         2.5   0  476   418
## 5 Samsung 970 Pro  21981         m.2   1 2327  2056
## 6 Samsung 860 Pro   2785         2.5   0  481   437
## 7 Samsung 850 Evo  92620         2.5   0  471   433
```

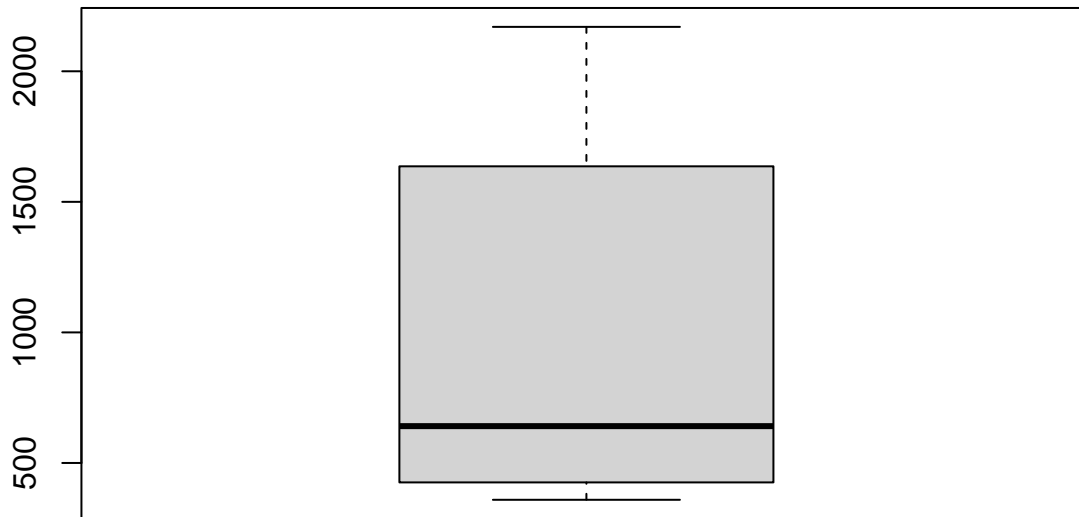
10. Which if any models blew away the competition in terms of their write speed? Show all work. There are different ways to go about solving this problem, and answers may vary.

```
ssd.speed.samsung[ order(ssd.speed.samsung$write, decreasing = TRUE), ]
```

```
##      brand  model samples form_factor nvme read write
## 8 Samsung 970 Evo Plus 101848         m.2   1 2381 2170
## 9 Samsung 970 Evo  121774         m.2   1 2342 2126
## 5 Samsung 970 Pro  21981         m.2   1 2327 2056
## 15 Samsung PM981  18364         m.2   1 1573 1710
```

```
## 3 Samsung      960 Evo  25990      m.2  1 1798 1562
## 16 Samsung     960 Pro  11872      m.2  1 2079 1462
## 14 Samsung      PM961   4374      m.2  1 1538 1280
## 12 Samsung      PM951   1938      m.2  1  968  641
## 10 Samsung     860 Evo 255164      2.5  0  485  444
##  6 Samsung     860 Pro   2785      2.5  0  481  437
##  7 Samsung     850 Evo  92620      2.5  0  471  433
##  4 Samsung     850 Pro  14690      2.5  0  476  418
## 11 Samsung     860 QVO  71901      2.5  0  478  400
##  2 Samsung     840 Evo  16888      2.5  0  470  389
## 13 Samsung     840 Evo    462     mSATA 0  469  359
```

```
boxplot(ssd.speed.samsung$write)
```



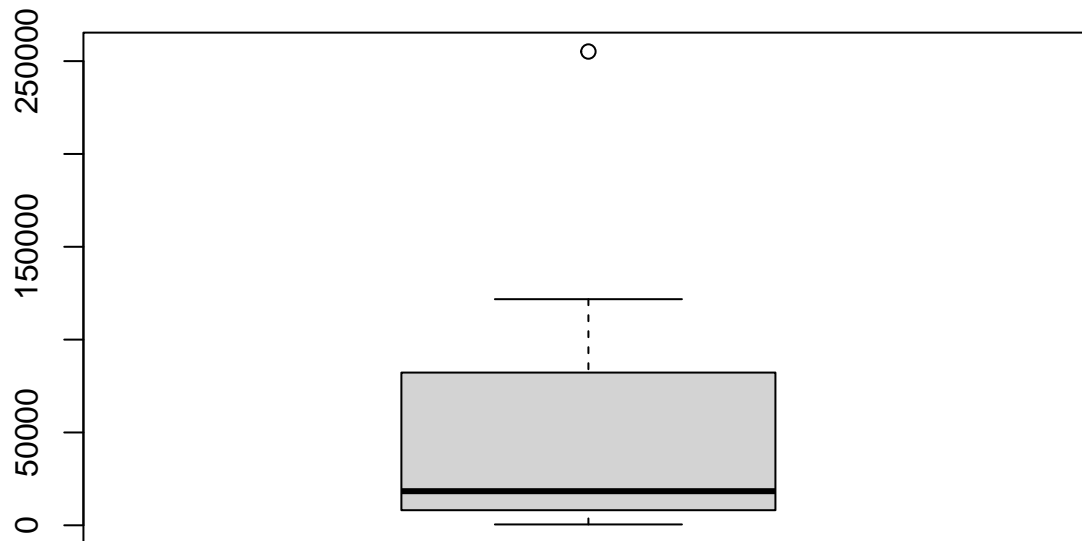
Doesn't seem like any models blew away the competition. There were obviously better and worse models, but no particular outliers, as shown in both the ordered data frame and boxplot.

11. Which if any models blew away the competition in terms of number of samples. Show all work. There are different ways to go about solving this problem.

```
ssd.speed.samsung[ order(ssd.speed.samsung$samples, decreasing = TRUE), ]
```

```
##      brand      model samples form_factor nvme read write
## 10 Samsung     860 Evo 255164         2.5   0  485  444
##  9 Samsung     970 Evo 121774         m.2   1 2342 2126
##  8 Samsung 970 Evo Plus 101848         m.2   1 2381 2170
##  7 Samsung     850 Evo  92620         2.5   0  471  433
## 11 Samsung     860 QVO  71901         2.5   0  478  400
##  3 Samsung     960 Evo  25990         m.2   1 1798 1562
##  5 Samsung     970 Pro  21981         m.2   1 2327 2056
## 15 Samsung      PM981  18364         m.2   1 1573 1710
##  2 Samsung     840 Evo  16888         2.5   0  470  389
##  4 Samsung     850 Pro  14690         2.5   0  476  418
## 16 Samsung     960 Pro  11872         m.2   1 2079 1462
## 14 Samsung      PM961   4374         m.2   1 1538 1280
##  6 Samsung     860 Pro   2785         2.5   0  481  437
## 12 Samsung      PM951   1938         m.2   1  968  641
## 13 Samsung     840 Evo    462        mSATA  0  469  359
```

```
boxplot(ssd.speed.samsung$samples)
```



860 Evo is an outlier as shown in the boxplot. Also twice as much as the next highest in the sorted data frame.