STATS 60 Summer 2020: HW1

1) Summarizing Data

Consider the following dataset

$$\{6, 8, 7, 9, 3, 7, 7, 10, 8, 2\}$$

1) What is the mean?

```
v = c(6,8,7,9,3,7,7,10,8,2)
mean(v)
```

```
## [1] 6.7
```

2) What is the median?

```
median(v)
```

```
## [1] 7
```

3) What is the mode?

The mode is 7.

4) What is the standard deviation?

Your answer depends on whether you interpreted this as a sample or population standard deviation (in this case either is fine). R calculates standard deviation using $N - 1$ in the denominator (you can check this using the command `?sd`).

```
s = sd(v) #sample sd
p = sd(v) * sqrt(9/10)
s
```

```
## [1] 2.496664
```

```
p
```

```
## [1] 2.368544
```

2) Introduction to R

1) Assign the character string "10" to `x` and "20" to `y`

```
x = '10'
y = '20'
```

2) What happens when you run `x + y`? Why does this happen?

```
try(x+y)
```

```
## Error in x + y : non-numeric argument to binary operator
```

We get an error because `x` and `y` are strings and R only takes in numerics as arguments to `+`.

3) How would we treat `x` and `y` as numbers to add them?

Either of these would work

```
x <- strtoi(x)
y <- strtoi(y)

##

x <- 10
y <- 10
```

4) Suppose I wanted to print the string "STATS60". I run the following code. Why do I get the following error? What should I do instead?

```
print(STATS60)
```

```
## Error in print(STATS60) : object 'STATS60' not found
```

Without the quotation marks, R interprets `STATS60` as an object. Since we have not initialized this object yet, R throws an error. To fix this, add quotation marks.

```
print('STATS60')
```

```
## [1] "STATS60"
```

Suppose now I wanted to print the string "060". I run the following code. Why do I get the incorrect output? What should I do instead?

```
print(060)
```

```
## [1] 60
```

Without quotation marks, R interprets `060` as a numeric. Thus, `060 = 60`, so R outputs 60. To fix this, add quotation marks.

```
print('060')
```

```
## [1] "060"
```

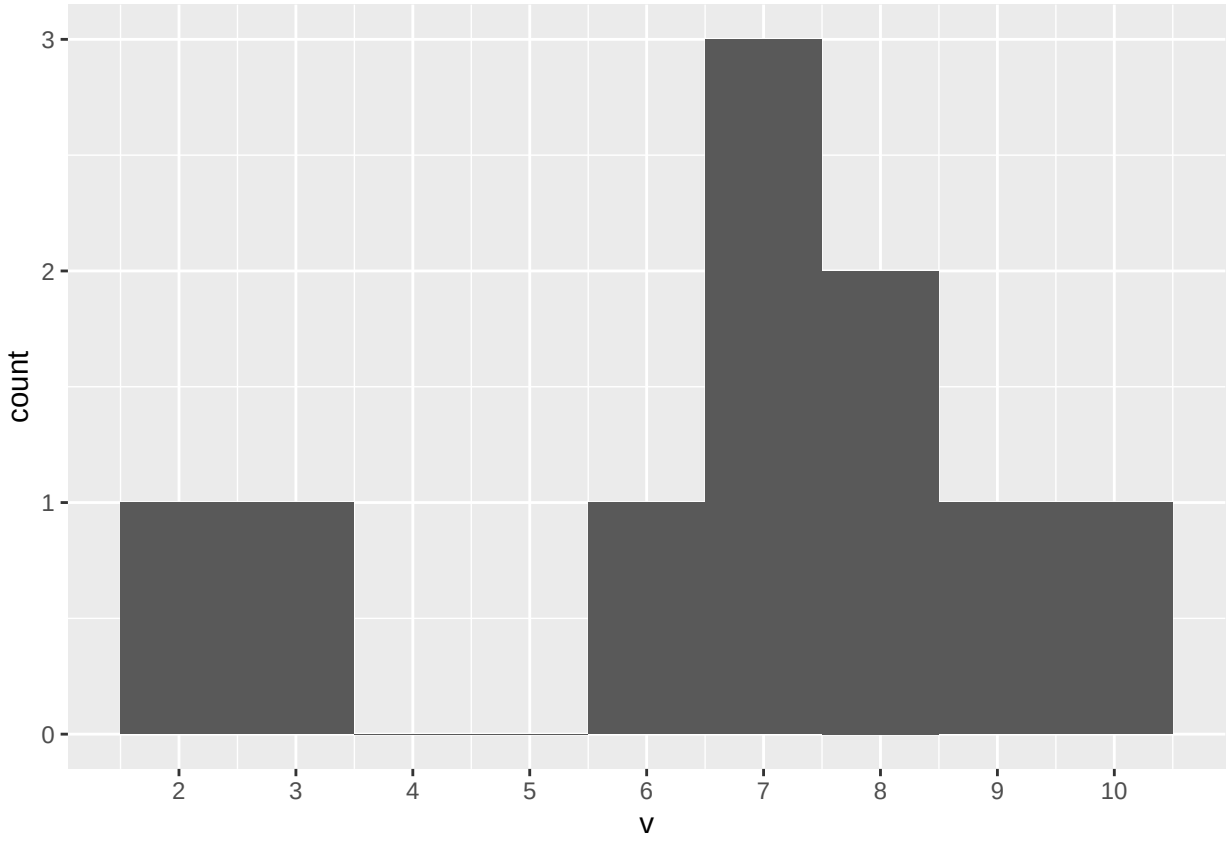
3) R for Data

1) Create a vector `v` of the data from Part 1.

```
v = c(6,8,7,9,3,7,7,10,8,2)
```

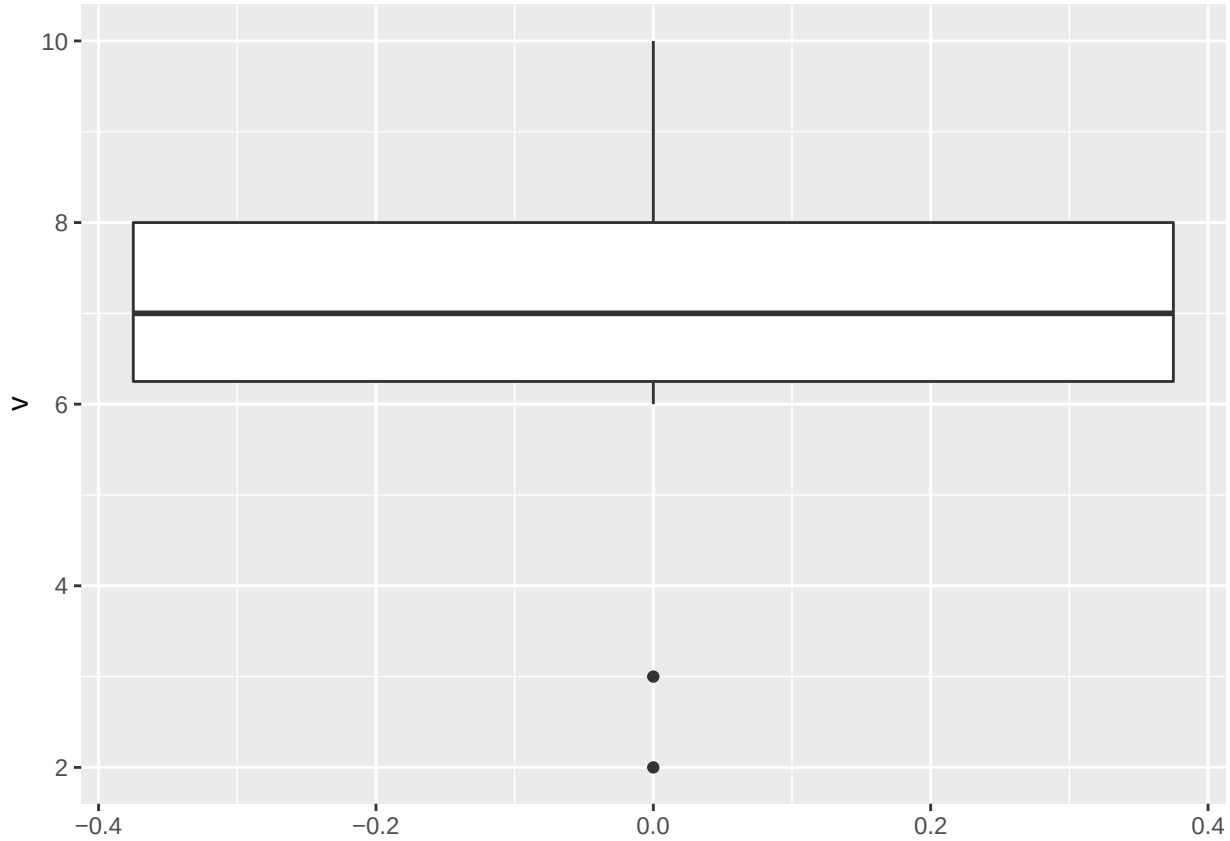
2) Create a histogram of `v`

```
library(ggplot2)
v = data.frame(v)
ggplot(v, aes(x = v)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = 1:12)
```



3) Create a boxplot of `v`. What is your interpretation of this?

```
ggplot(v, aes(y = v)) +
  geom_boxplot()
```



The median is 7 and the 25th quartile is 6.25 and the 75th quartile is 8. Because of this IQR, 2 and 3 are outliers. It seems like most of the data is concentrated around 7 otherwise.

4) Write code that gives you the mean and standard deviation of `v`.

We already did this above, but

```
sd(v$v)
```

```
## [1] 2.496664
```