Class19

R(PID:A59010419

12/1/2021

```
data <- read.csv('373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv')
```

```
table(data$Genotype..forward.strand.)
```

```
##
## A|A A|G G|A G|G
##  22  21  12   9
```

```
table(data$Genotype..forward.strand.)/nrow(data)
```

```
##
##      A|A      A|G      G|A      G|G
## 0.343750 0.328125 0.187500 0.140625
```

```
table1 <- read.table('rs8067378_ENSG00000172057.6.txt')
```

Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
summary(table1)
```

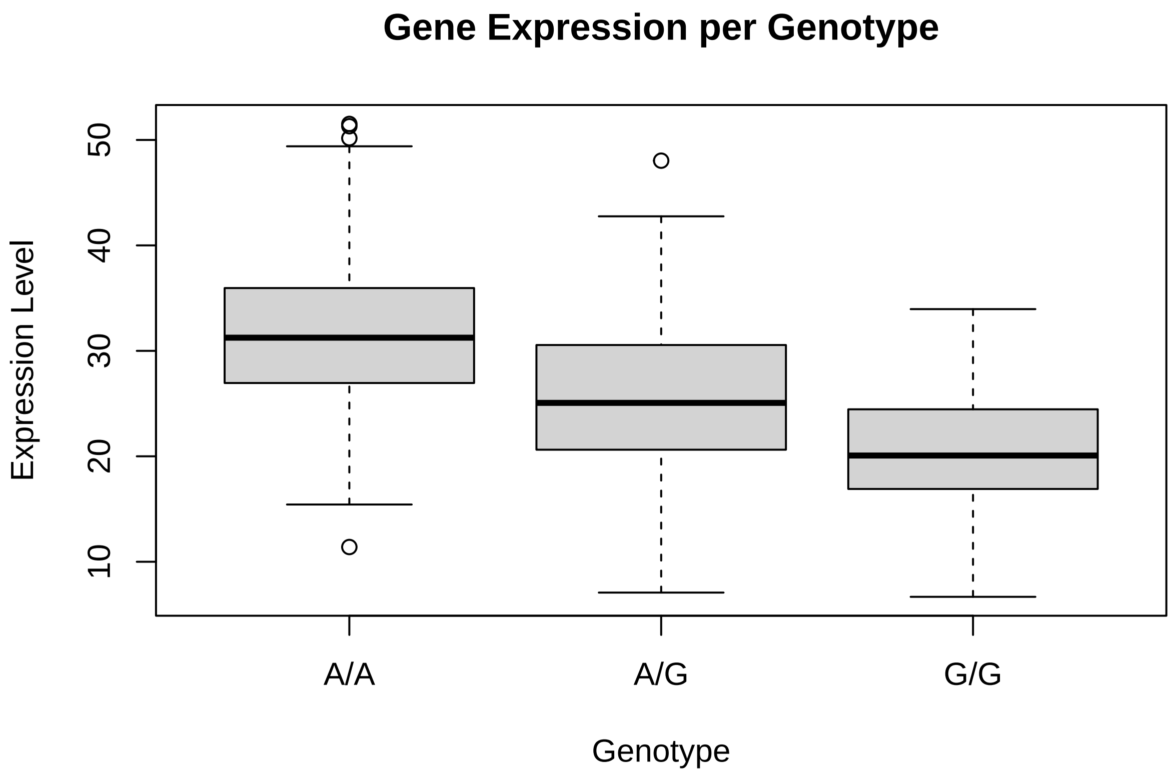
```
##     sample              geno                exp
##  Length:462         Length:462         Min.   : 6.675
##  Class :character   Class :character   1st Qu.:20.004
##  Mode  :character   Mode  :character   Median :25.116
##                                        Mean   :25.640
##                                        3rd Qu.:30.779
##                                        Max.   :51.518
```

There are 462 genotypes in the data talbe.

```
table(table1$geno)
```

```
##
## A/A A/G G/G
## 108 233 121
```

```r
boxplot1 <- boxplot(table1$exp~table1$geno,data=table1, main="Gene Expression per Genotype",
   xlab="Genotype", ylab="Expression Level")
```



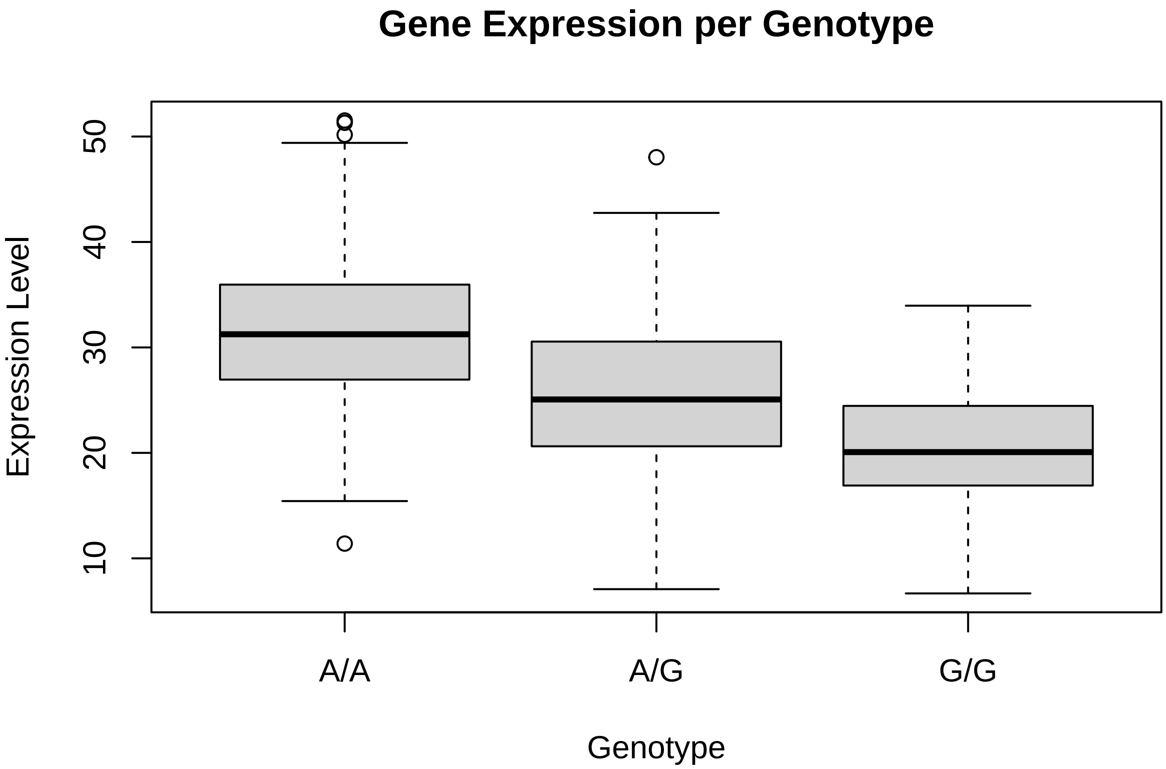
```r
boxplot1
```

```
## $stats
##          [,1]     [,2]     [,3]
## [1,] 15.42908  7.07505  6.67482
## [2,] 26.95022 20.62572 16.90256
## [3,] 31.24847 25.06486 20.07363
## [4,] 35.95503 30.55183 24.45672
## [5,] 49.39612 42.75662 33.95602
##
## $n
## [1] 108 233 121
##
## $conf
##          [,1]     [,2]     [,3]
## [1,] 29.87942 24.03742 18.98858
## [2,] 32.61753 26.09230 21.15868
##
## $out
## [1] 51.51787 50.16704 51.30170 11.39643 48.03410
##
## $group
## [1] 1 1 1 1 2
##
## $names
## [1] "A/A" "A/G" "G/G"
```

Based off the stats of the boxplot1 object the medians expression levels are 31.24847 25.06486 20.07363 for A|A,A|G, and G|G respectively. And from the table(table1geno) function and the n we found the sample size to be 108 233 121 for A|A,A|G, and G|G respectively.

Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3? Repeat of above table for description of answer.

```
boxplot1 <- boxplot(table1$exp~table1$geno,data=table1, main="Gene Expression per Genotype",
   xlab="Genotype", ylab="Expression Level")
```



A|A 15.42908 26.95022 31.24847 35.95503 49.39612 G|G 6.67482 16.90256 20.07363 24.45672 33.95602

And above I have taken the values from the Stats for the AA and GG genotype.

At first glance the AA SNP seems to increase expression levels compared to the GG SNP. This can be seen as an increase in median. Further examination we see the lower quartile for AA is 26.9 and the upper quartile for GG is 24.45. This non overlapping quartile could be taken to say that there is in fact an increase in AA expression compared to GG expression. This would be based on the claim that the middle 50% of the data for AA has higher expression than the middle 50% of the data for GG. Further statistical testing would need to be done to make any further claims. It is clear that some of the data in these two distributions overlap so it may not be a significant difference from other statistical approaches.