Homework 1

Matthew Nickols Problem 1

```
su = read.delim("DATA/Su_raw_matrix.txt")
Liver_mean = mean(su$Liver_2.CEL)
Liver_sd = sd(su$Liver_2.CEL)
column_means = colMeans(su)
column_sums = colSums(su)

Liver_mean
```

```
## [1] 241.8246
```

```
Liver_sd
```

```
## [1] 1133.352
```

```
column_means
```

```
##       Brain_1.CEL       Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL
##          204.9763          315.0924          198.3439          267.6551
## Fetal_liver_1.CEL Fetal_liver_2.CEL       Liver_1.CEL       Liver_2.CEL
##          209.8722          399.1482          160.8558          241.8246
```

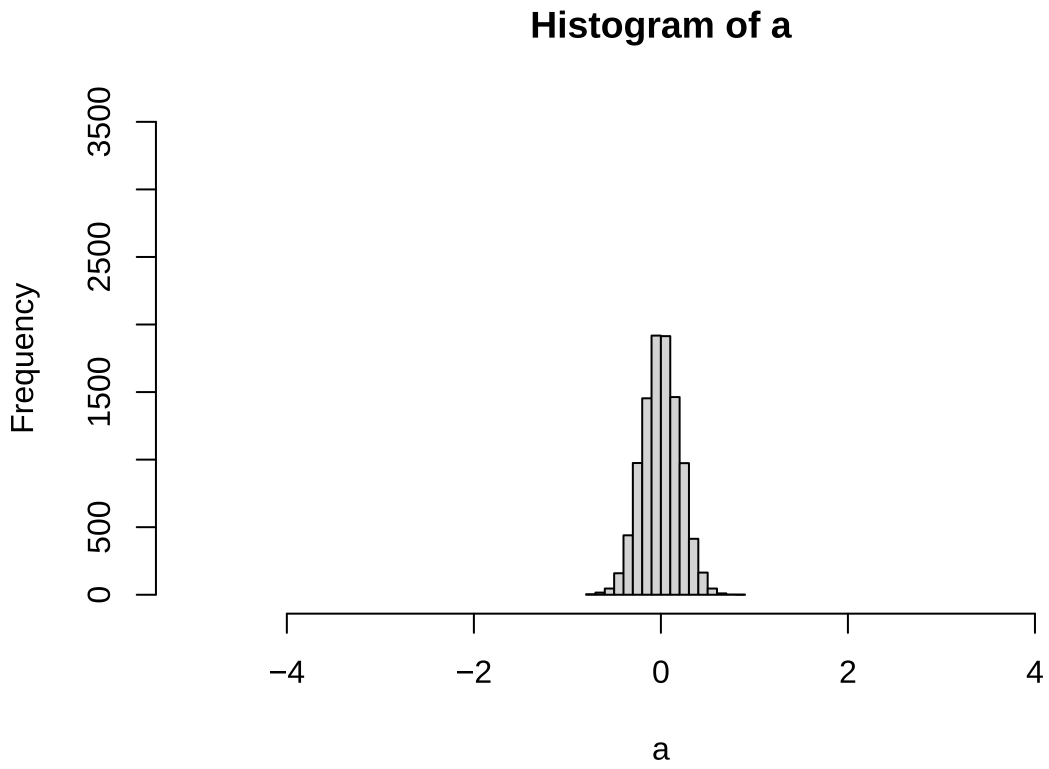
```
column_sums
```

```
##       Brain_1.CEL       Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL
##           2588031           3978357           2504290           3379413
## Fetal_liver_1.CEL Fetal_liver_2.CEL       Liver_1.CEL       Liver_2.CEL
##           2649846           5039645           2030966           3053278
```

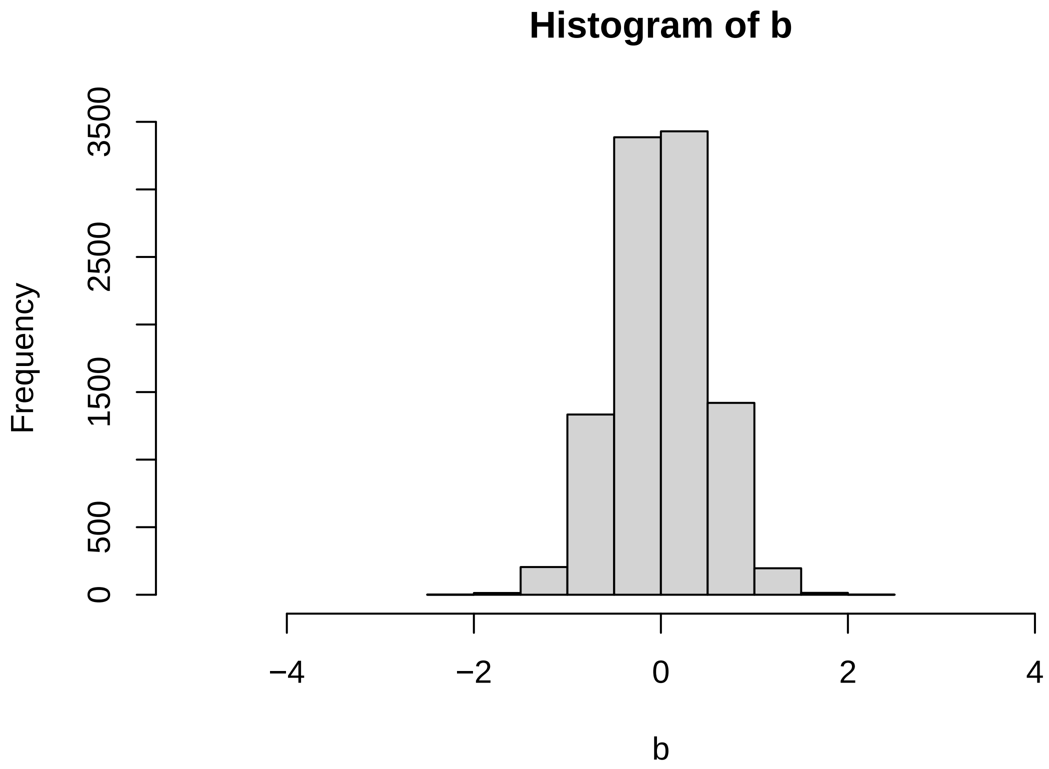
Problem 2

```
a = rnorm(10000, 0, 0.2)
b = rnorm(10000, 0, 0.5)

hist(a, xlim=c(-5,5), ylim=c(0,3500))
```



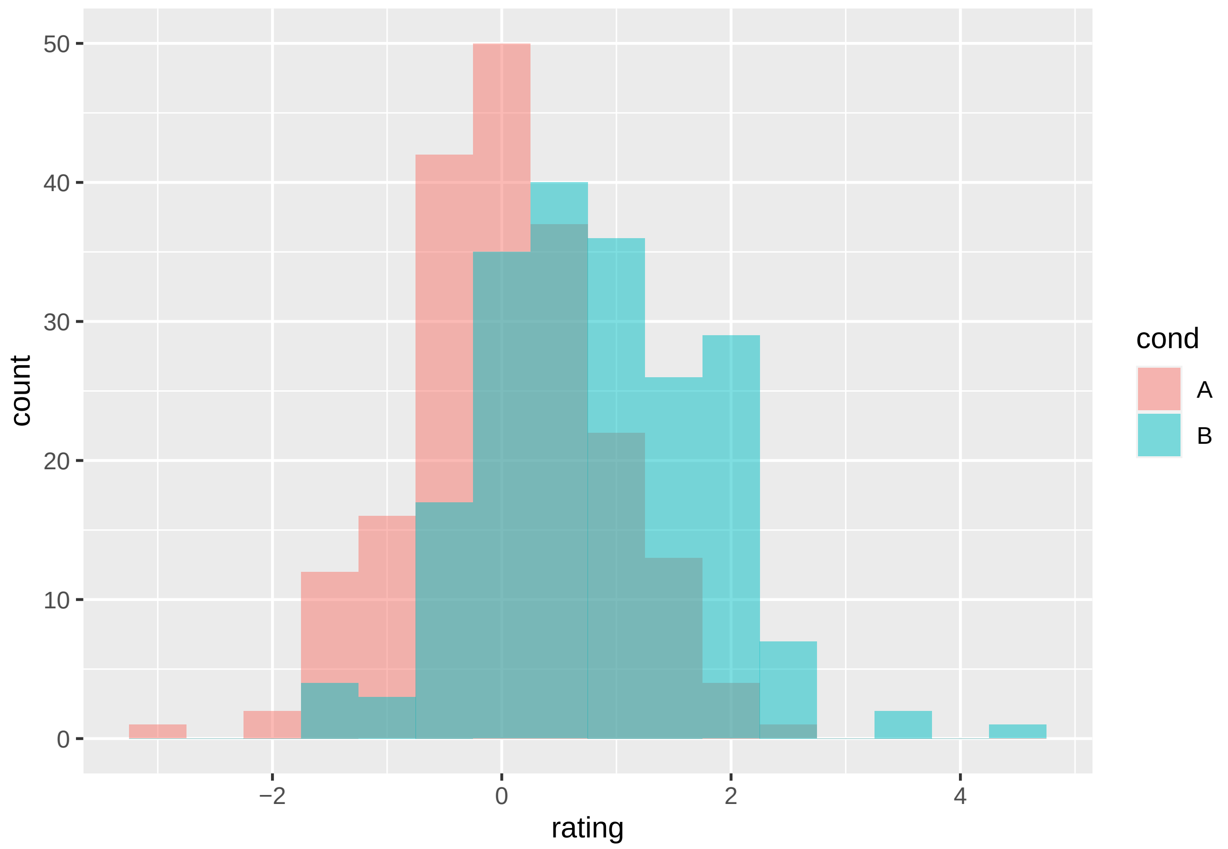
```
hist(b, xlim=c(-5,5), ylim=c(0,3500))
```



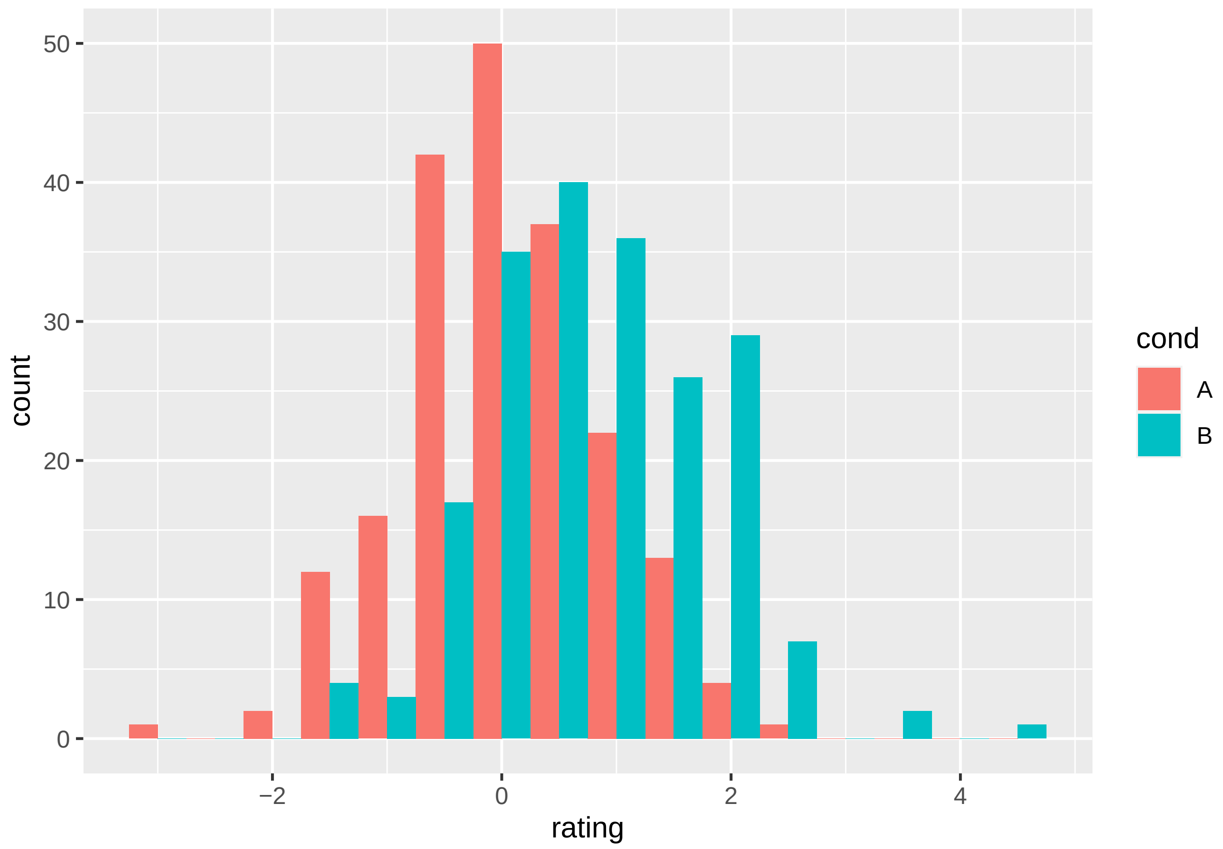
These histograms are different in the range of x-values that they have. Histogram a has less variance, with values ranging from -1 to 1, with the majority being between -0.5 to 0.5. Histogram b has x values from -2 to 2 with most of them being between -1 and 1. This means that there is a larger variance to histogram b, and histogram a has a larger frequency.

Problem 3

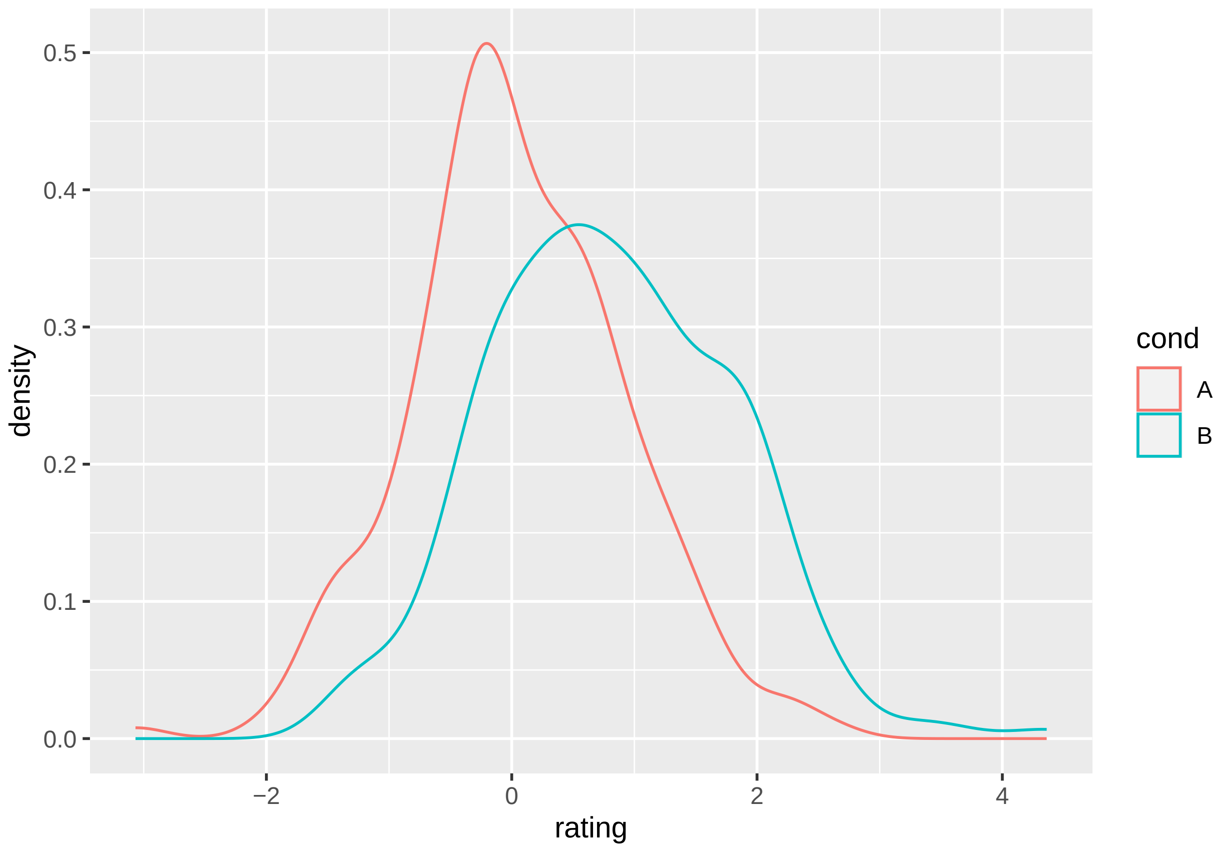
```
library(ggplot2)
dat <- data.frame(cond = factor(rep(c("A","B"), each=200)), rating = c(rnorm(200), rnorm(200, mean=.8)))
# Overlaid histograms
ggplot(dat, aes(x=rating, fill=cond)) +
geom_histogram(binwidth=.5, alpha=.5, position="identity")
```



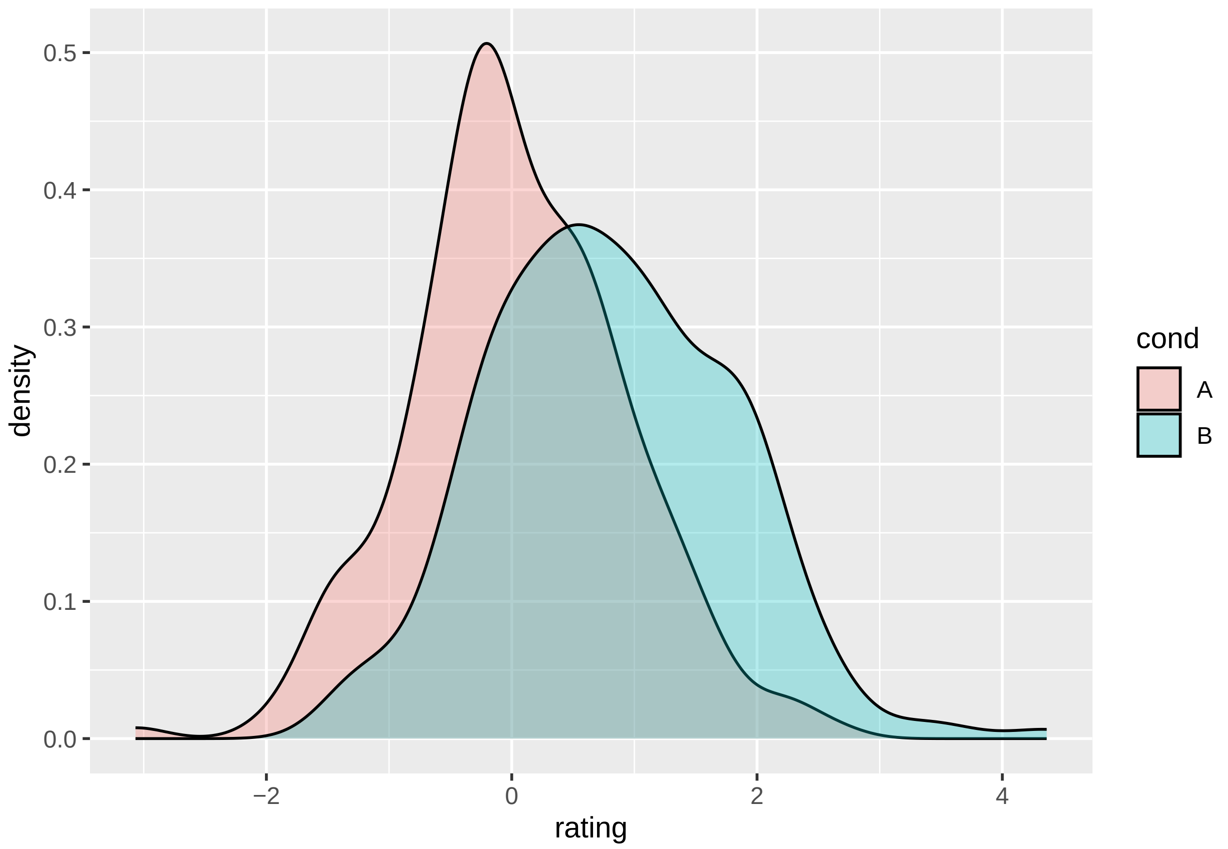
```
# Interleaved histograms
ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5, position="dodge")
```



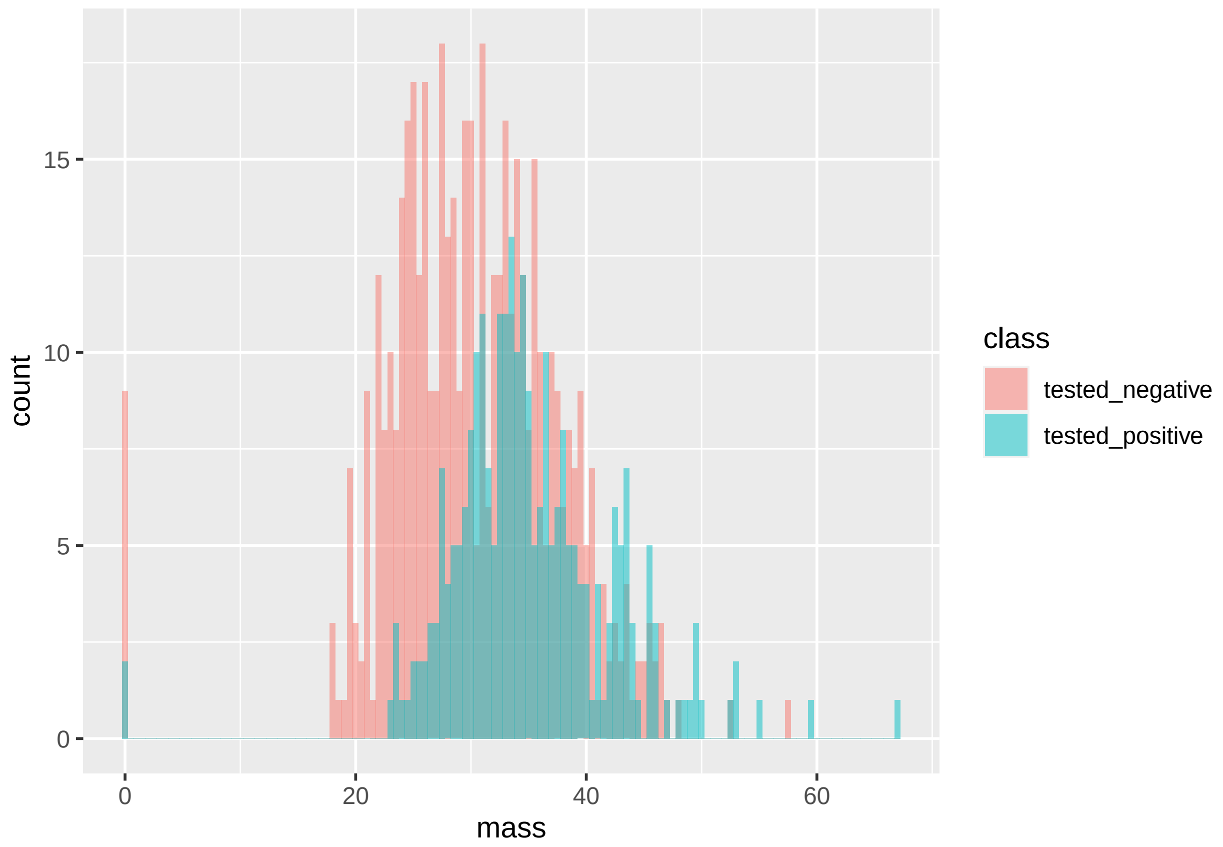
```
# Density plots
ggplot(dat, aes(x=rating, colour=cond)) + geom_density()
```



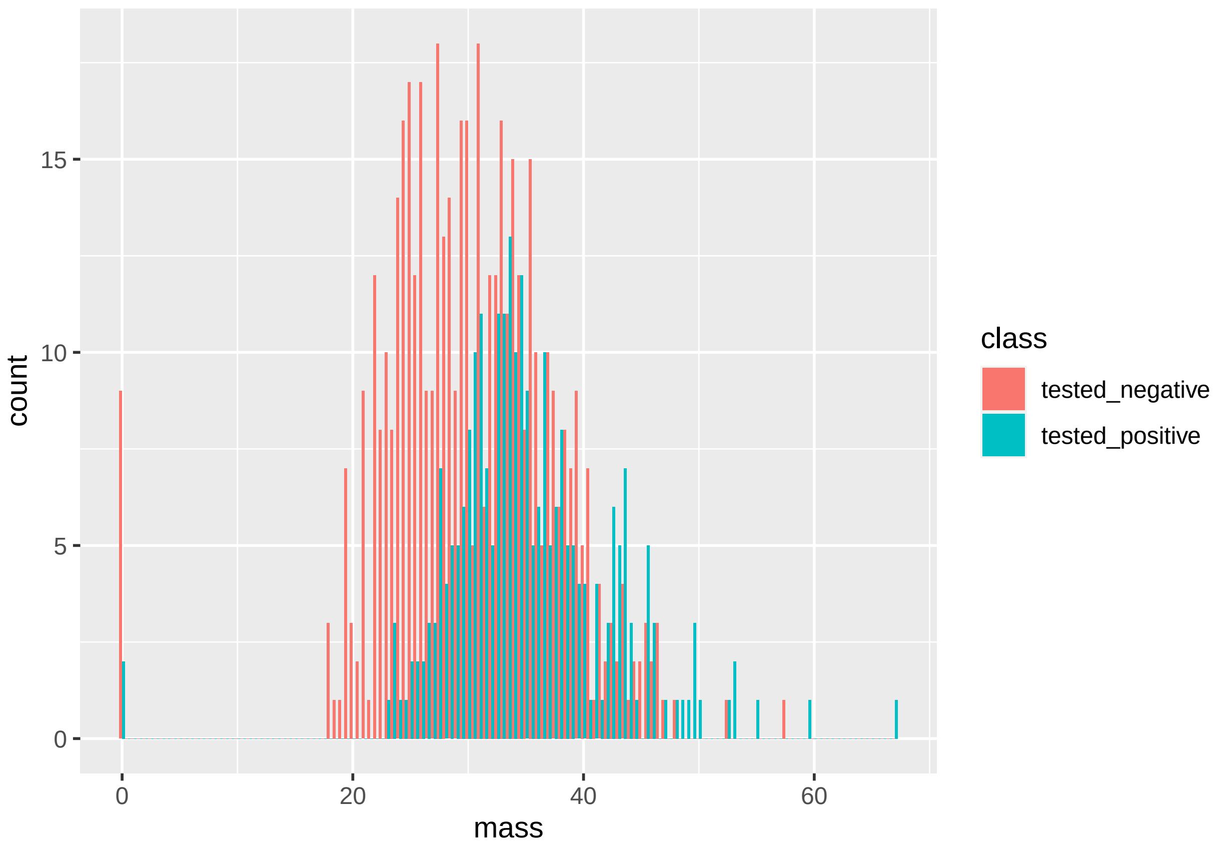
```
# Density plots with semitransparent fill
ggplot(dat, aes(x=rating, fill=cond)) + geom_density(alpha=.3)
```



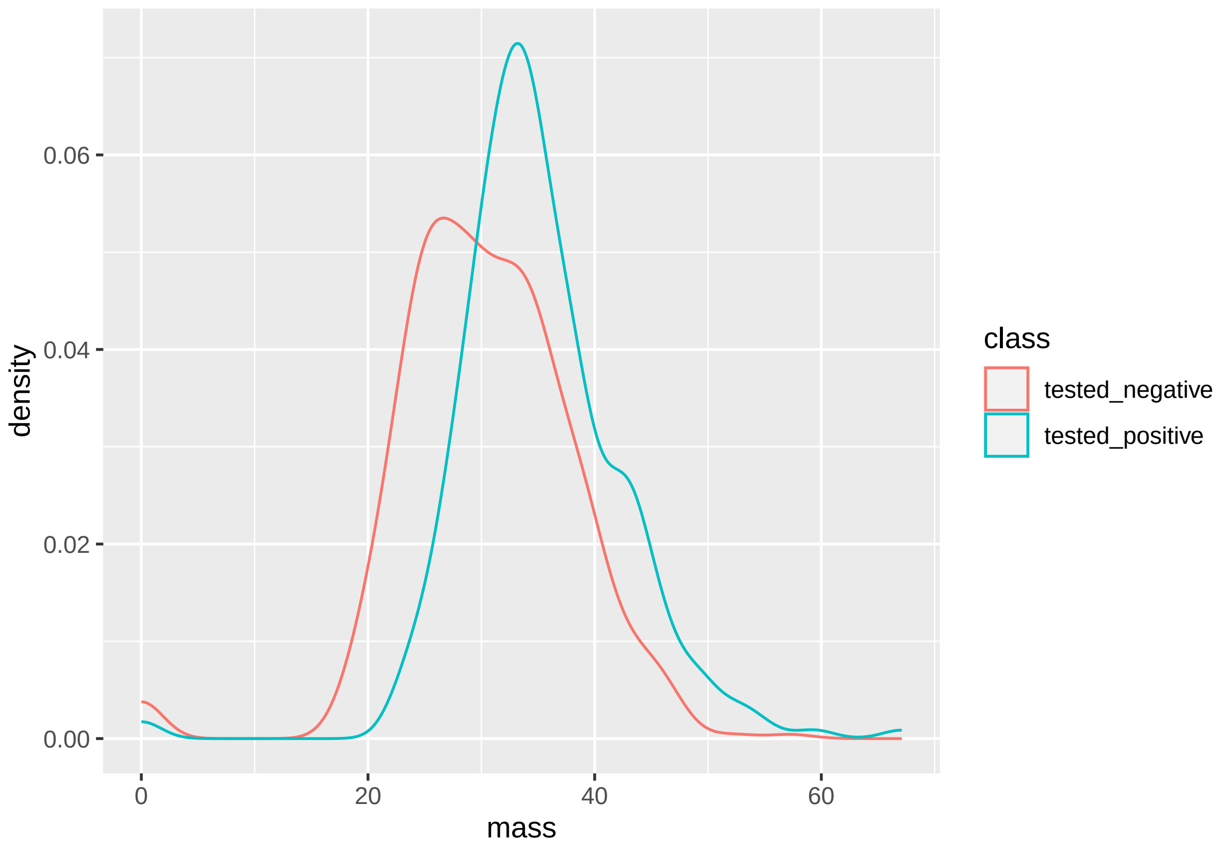
```
# Reading diabetes_train.csv and creating graphs
diabetes <- read.csv("DATA/diabetes_train.csv")
# Overlaid histogram
ggplot(diabetes, aes(x=mass, fill=class)) +
geom_histogram(binwidth=.5, alpha=.5, position="identity")
```



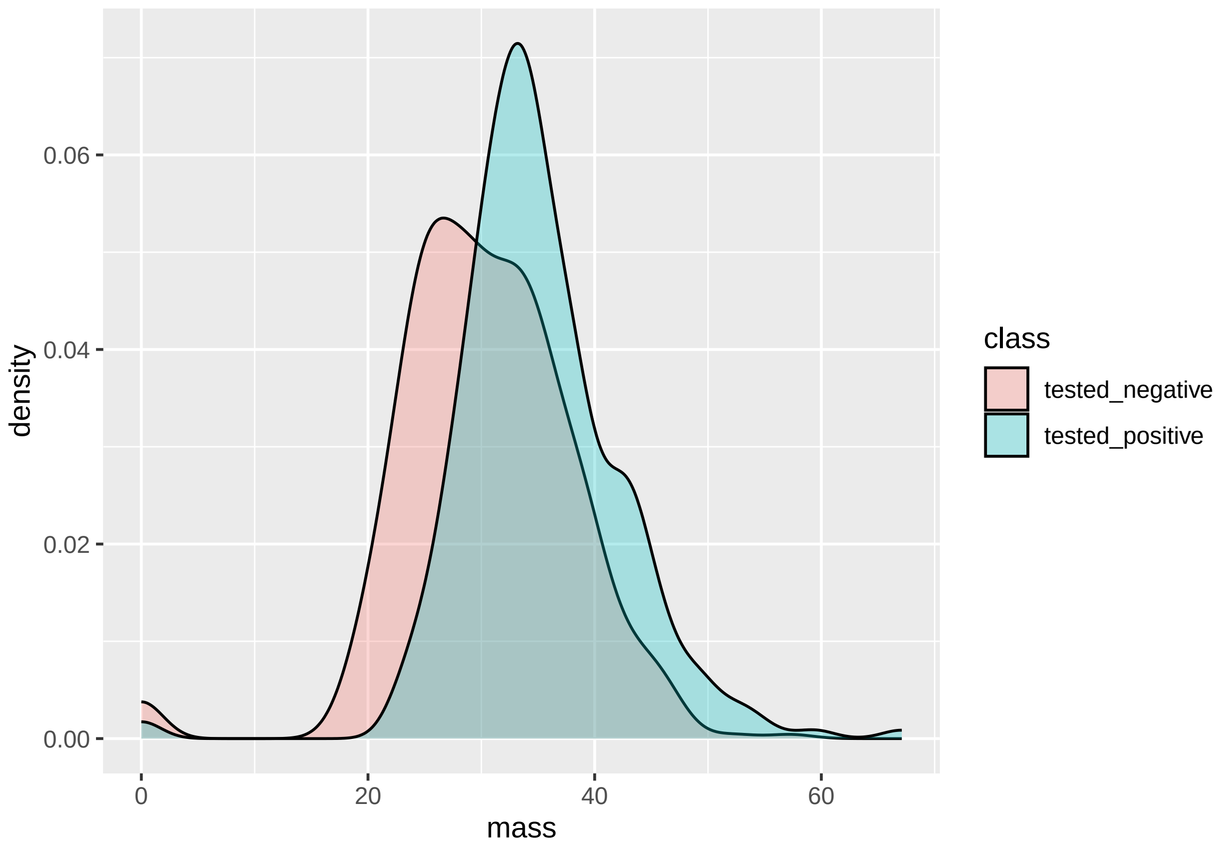
```
# Interleaved histograms
ggplot(diabetes, aes(x=mass, fill=class)) + geom_histogram(binwidth=.5, position="dodge")
```



```r
# Density plots
ggplot(diabetes, aes(x=mass, colour=class)) + geom_density()
```



```
# Density plots with semitransparent fill
ggplot(diabetes, aes(x=mass, fill=class)) + geom_density(alpha=.3)
```



Problem 4

```
# Reading in csv file
passengers <- read.csv("DATA/titanic.csv")

#Commenting out these lines so they do not output when knitting
#First operation drops rows that have a missing value (NA) and outputs a summary of each column of data
#passengers %>% drop_na() %>% summary()

#Second operation shows us all rows of male passengers
#passengers %>% filter(Sex == "male")

#Third operation arranges them with the highest paid Fare at the top and the lowest paid Fare at the bo
#passengers %>% arrange(desc(Fare))

#Fourth operation creates a new column called FamSize and populates it with Parch + SibSp
#passengers %>% mutate(FamSize = Parch + SibSp)

#Last operation splits the passengers into male or female and then gives the average Fare paid by eithe
#passengers %>% group_by(Sex) %>% summarise(meanFare = mean(Fare), numSurv = sum(Survived))
```

Problem 5

```
# Reading diabetes_train.csv and creating graphs
diabetes <- read.csv("DATA/diabetes_train.csv")

quantile(diabetes$skin, c(.10, .30, .5, .6))
```

```
## 10% 30% 50% 60%
##   0  10  23  27
```