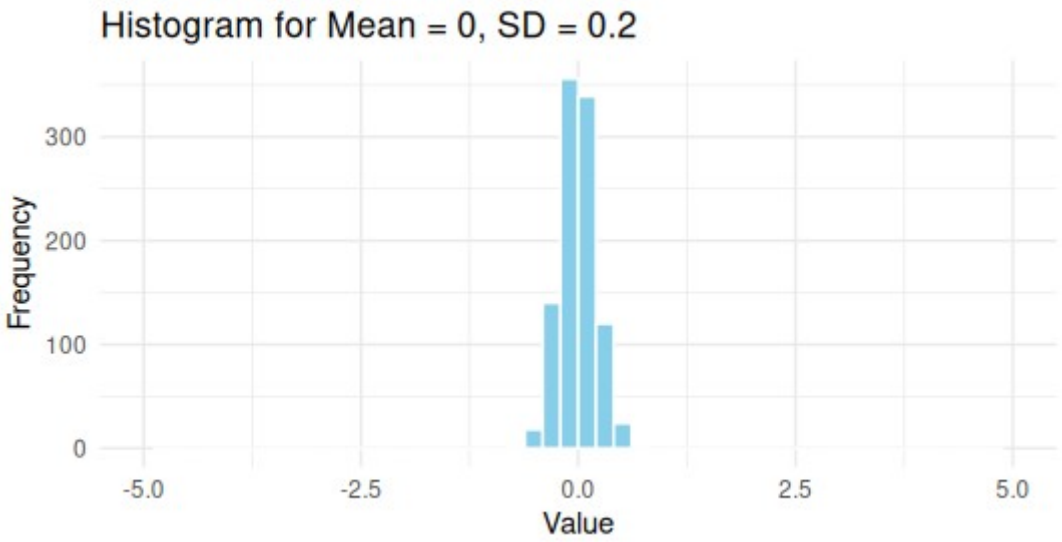
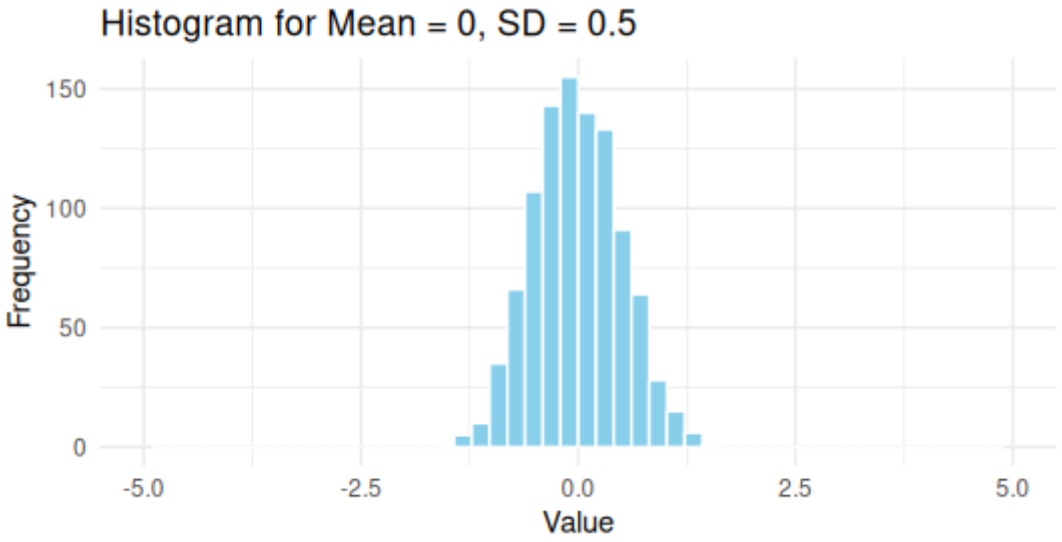
```
# 1. Use ”Su_raw_matrix.txt” for the following questions (30 points).
# Preliminary Stuff
install.packages('systemfonts')
install.packages('textshaping')

install.packages(c('googledrive', 'googlesheets4', 'httr', 'ragg', 'rvest', 'xml2'))
install.packages("ggplot2")
install.packages("tidyverse")

library(ggplot2)

# Problem 1
# Load the file
su <- read.delim('/home/jj/Downloads/Su_raw_matrix.txt')

# Mean
msu <- mean(su$Liver_2.CEL)

# Standard deviation
sdsu <- sd(su$Liver_2.CEL)

# Col Means
cmsu <- colMeans(su)

# Col Sums
cssu <- colSums(su)

# Problem 2
# Random 1000 matrices
randNumPoint2 <- rnorm(1000, mean = 0, sd = .2)
randNumPoint5 <- rnorm(1000, mean = 0, sd = .5)
df2 <- data.frame(values = randNumPoint2)
df3 <- data.frame(values = randNumPoint5)

# The .2 Histogram
ggplot(df2, aes(x = values)) +
  xlim(c(-5,5)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "white") +
  labs(title = "Histogram for Mean = 0, SD = 0.2",
       x = "Value",
       y = "Frequency") +
  theme_minimal()
```

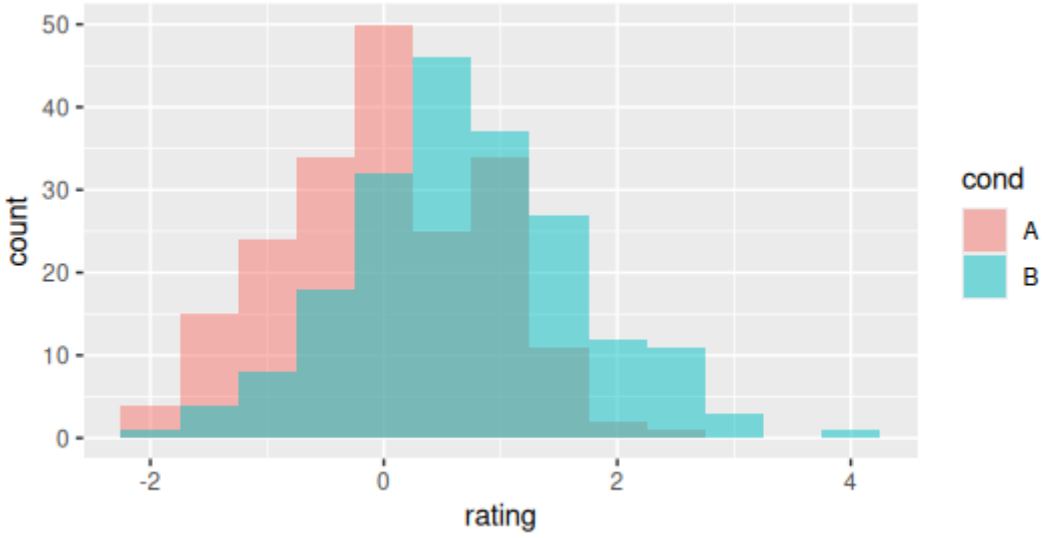



```
df3 <- data.frame(values = randNumPoint5)

# The .5 Histogram
ggplot(df3, aes(x = values)) +
  xlim(c(-5,5)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "white") +
  labs(title = "Histogram for Mean = 0, SD = 0.5",
       x = "Value",
       y = "Frequency") +
  theme_minimal
```

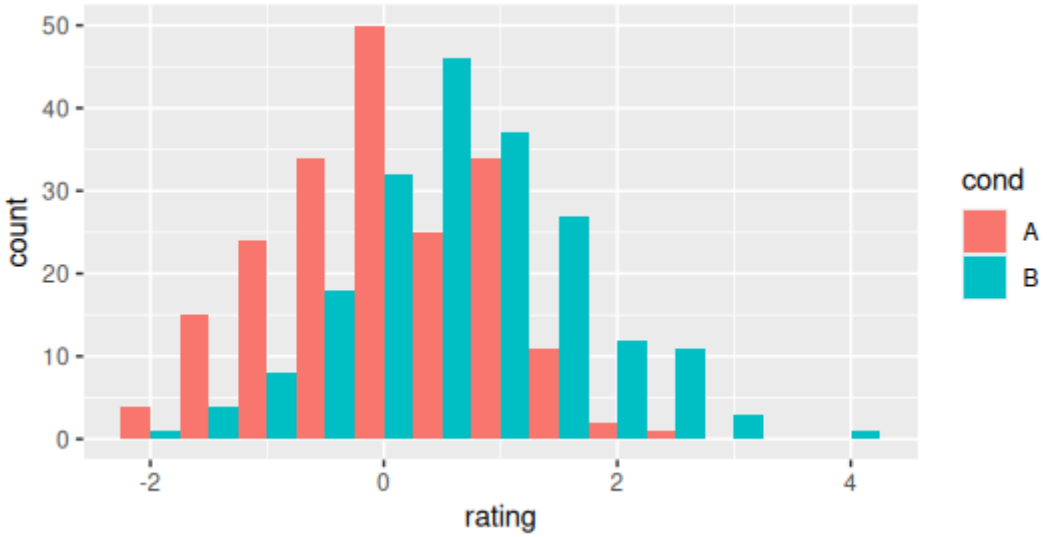



```
# Then comment on how these histograms are different from each other and state the reason.
# This essentially means a shorter, wider histogram
# The change in standard deviation from 0.2 to 0.5 results in
# a wider distribution over values that deviate more from the mean.

# Problem 3
# a) The setup
dat <- data.frame(cond = factor(rep(c("A","B"), each=200)),
```

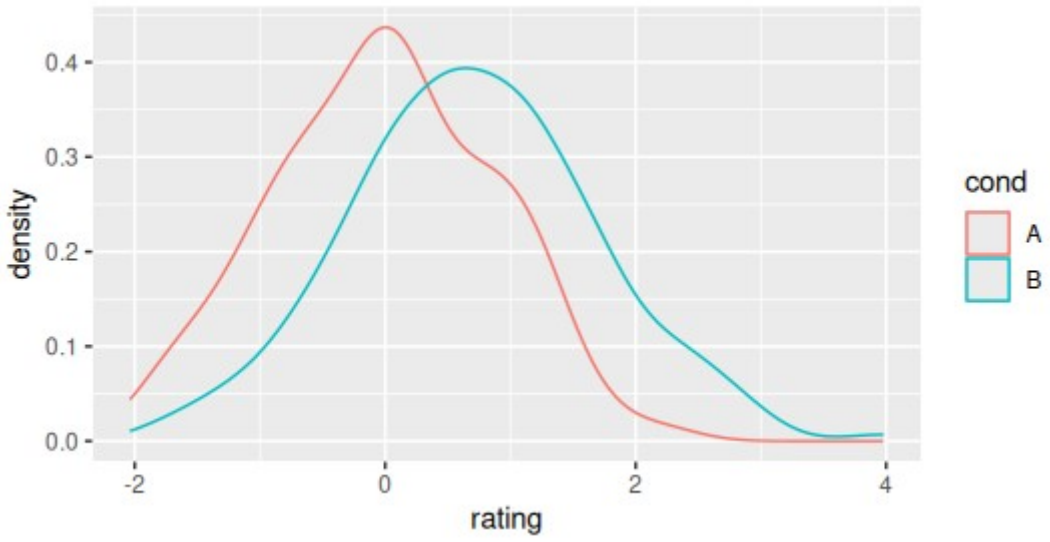
```
                rating = c(rnorm(200),rnorm(200, mean=.8)))
# b) Overlaid histograms
ggplot(dat, aes(x=rating, fill=cond)) +
  geom_histogram(binwidth=.5, alpha=.5, position="identity")
```



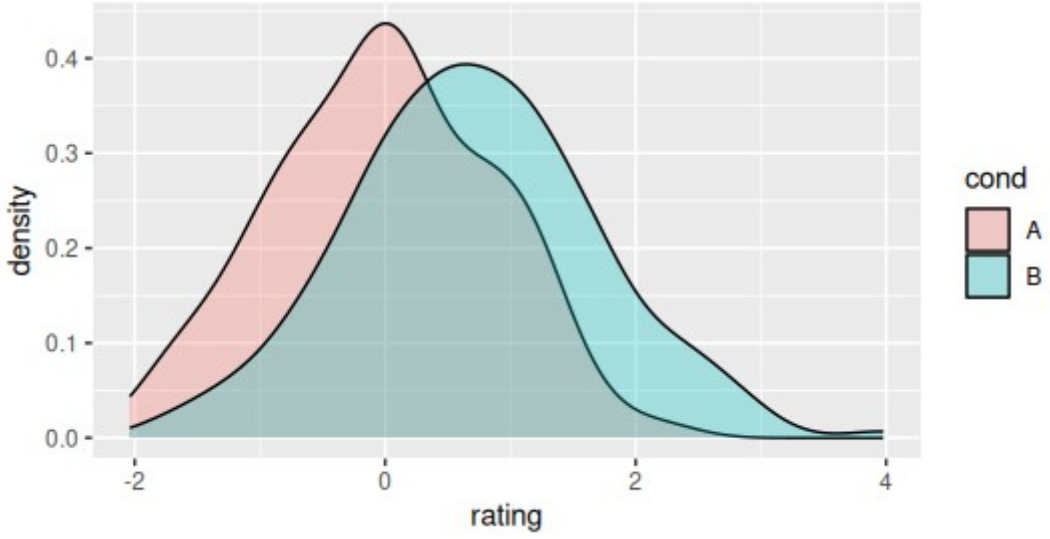

```
# c) Interleaved histograms
ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5, position="dodge")
```



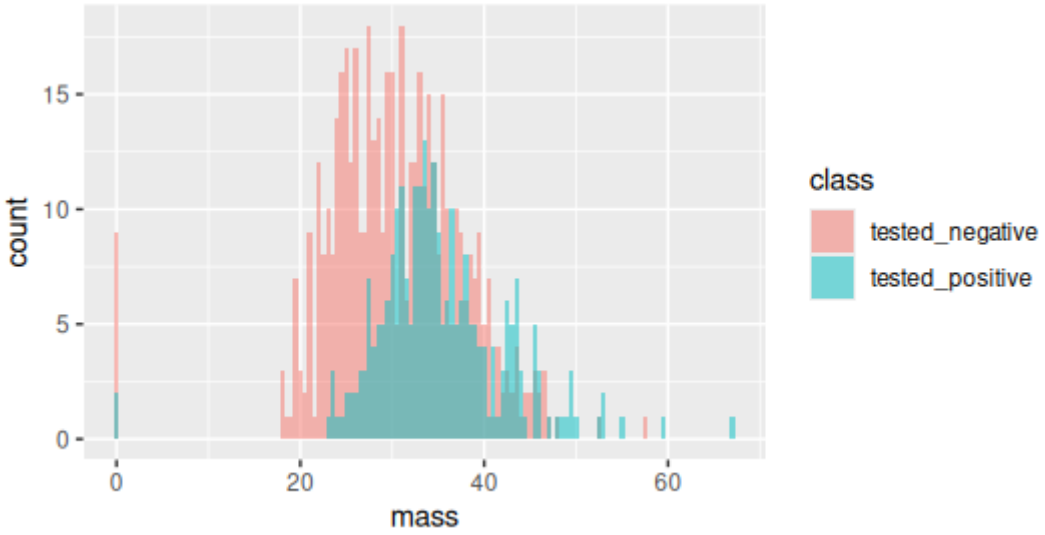

```
# d) Density plots
ggplot(dat, aes(x=rating, colour=cond)) + geom_density()
```



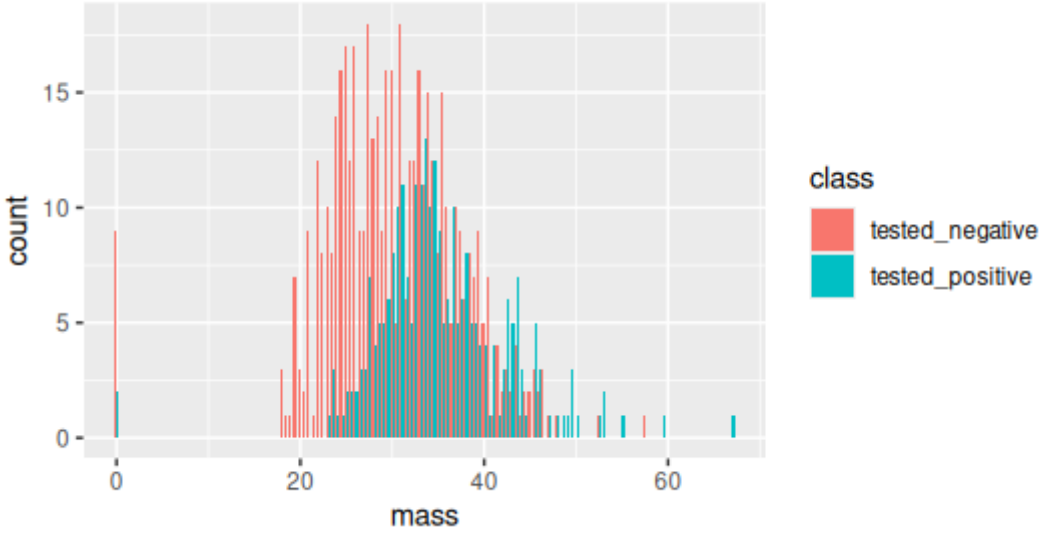
# e)Density plots with semitransparent fill
ggplot(dat, aes(x=rating, fill=cond)) + geom_density(alpha=.3)



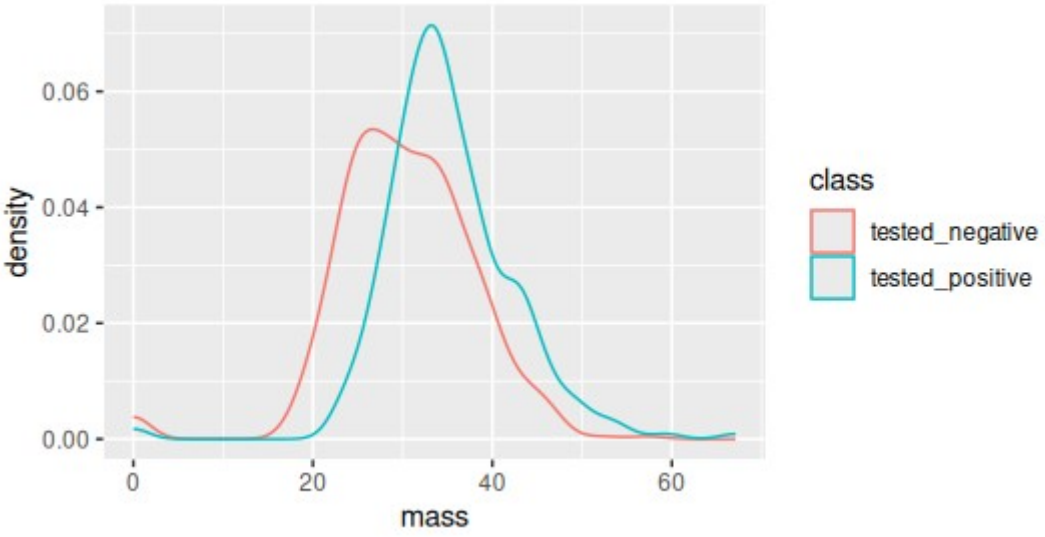
# f-a)
diabetes <- read.csv('/home/jj/Downloads/diabetes_train.csv')
# f-b) Overlaid histograms
ggplot(diabetes, aes(x=mass, fill=class)) +
  geom_histogram(binwidth=.5, alpha=.5, position="identity")



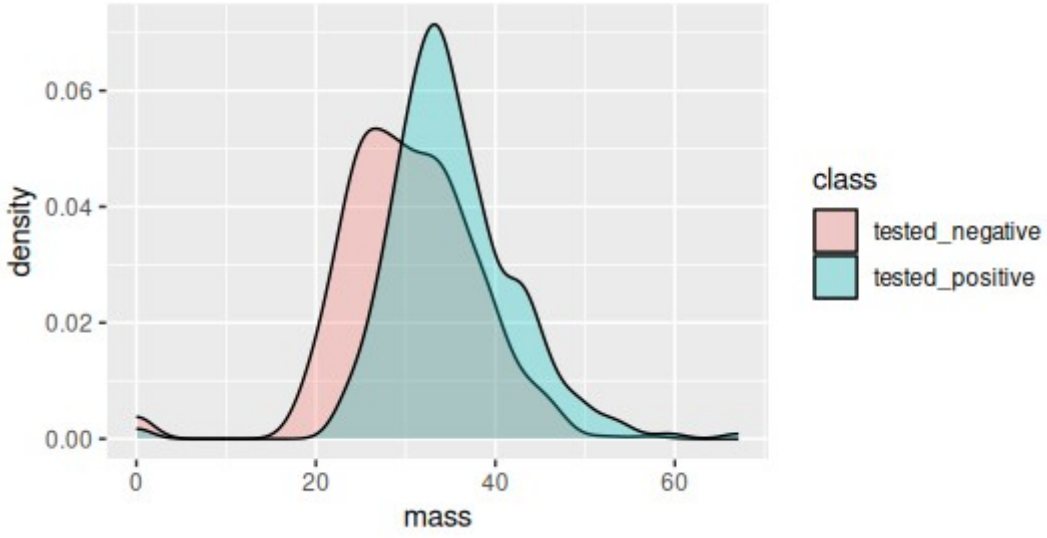
# f-c) Interleaved histograms
ggplot(diabetes, aes(x=mass, fill=class)) + geom_histogram(binwidth=.5, position="dodge")



# f-d) Density plots
ggplot(diabetes, aes(x=mass, colour=class)) + geom_density()



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



# Problem 4
library(tidyverse)
passengers <- read.csv('/home/jj/Downloads/Data/titanic.csv')

(a) Drop all non-numerical or null values from the dataset and print a summary
passengers %>% drop_na() %>% summary()
(b) Return only the rows where the Sex variable='male'
passengers %>% filter(Sex == "male")
(c) Sort the dataset by the fare column
passengers %>% arrange(desc(Fare))
(d) Add a column named family size that is the sum of the columns Parch and SibSp
passengers %>% mutate(FamSize = Parch + SibSp)
#(e) Get the average fare and add up number of survivors by sex
passengers %>% group_by(Sex) %>% summarise(meanFare = mean(Fare), numSurv = sum(Survived))

# Problem 5
quantiles <- quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60))