

# HW1

Mason Turner

1/23/2024

Note: This assumes my working directory is C:\Users\mctur\OneDrive\Desktop\Advanced\_Data\_Mining\datamining-main\Rscripts

**Importing necessary packages.**

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v lubridate  1.9.3      v tibble    3.2.1
## v purrr      1.0.2      v tidyr     1.3.0
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(knitr)
```

**Question 1.a - Reading Su\_raw\_matrix.txt into a variable called 'su' using the 'read.delim' function.**

```
filename = 'data/Su_raw_matrix.txt'
su <- read.delim(filename, header = TRUE)
```

**Question 1.b - Using 'mean' and 'sd' functions to find mean and standard deviation of Liver\_2.CEL column. Then printing them out.**

```
L2C_mean <- mean(su$Liver_2.CEL)
L2C_sd <- sd(su$Liver_2.CEL)
L2C_mean
```

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

```
L2C_sd
```

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

Question 1.c - Using the 'colMeans' and 'colSums' functions to get the average and total values of each column. Then printing them out.

```
mean_values <- colMeans(su)
sum_values <- colSums(su)
mean_values
```

```
##      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
```

```
sum_values
```

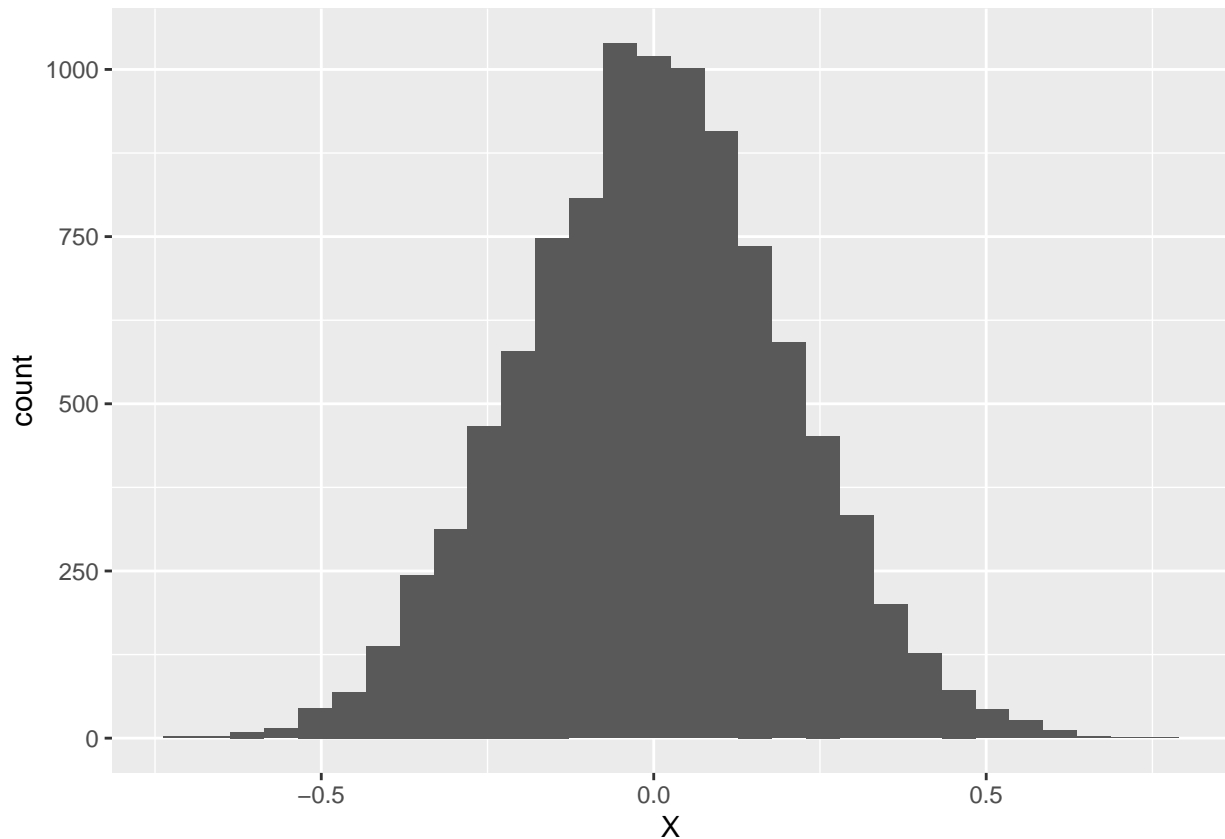
```
##      Brain_1.CEL      Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL
##      2588031      3978357      2504290      3379413
## Fetal_liver_1.CEL Fetal_liver_2.CEL      Liver_1.CEL      Liver_2.CEL
##      2649846      5039645      2030966      3053278
```

Question 2 - Generating 10000 numbers for the following (mean, sigma) pairs and plotting histograms for each.

Question 2.a - mean=0, sigma=0.2

```
ggplot(data.frame(X = rnorm(10000, 0, 0.2)), aes(x = X)) + geom_histogram()
```

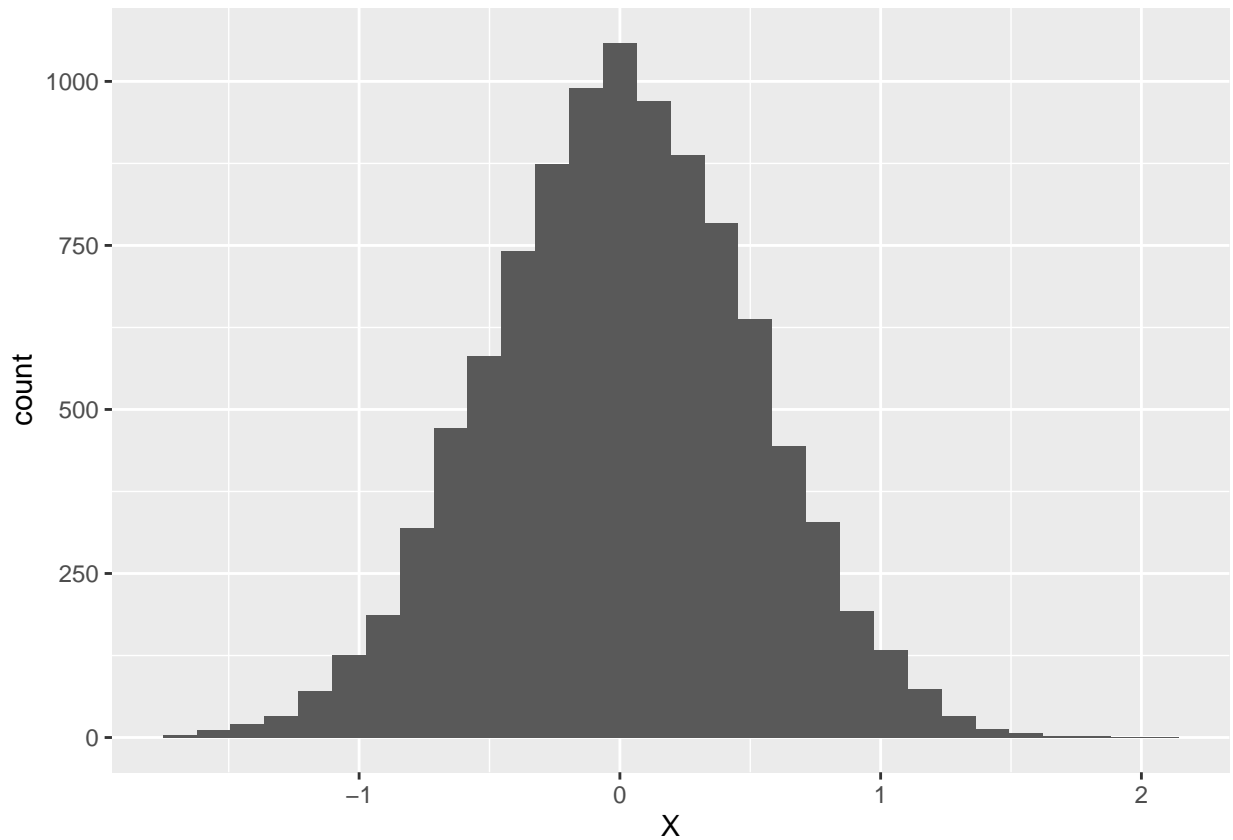
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Question 2.b - mean=0, sigma=0.5

```
ggplot(data.frame(X = rnorm(10000, 0, 0.5)), aes(x = X)) + geom_histogram()
```

## 'stat\_bin()' using 'bins = 30'. Pick better value with 'binwidth'.



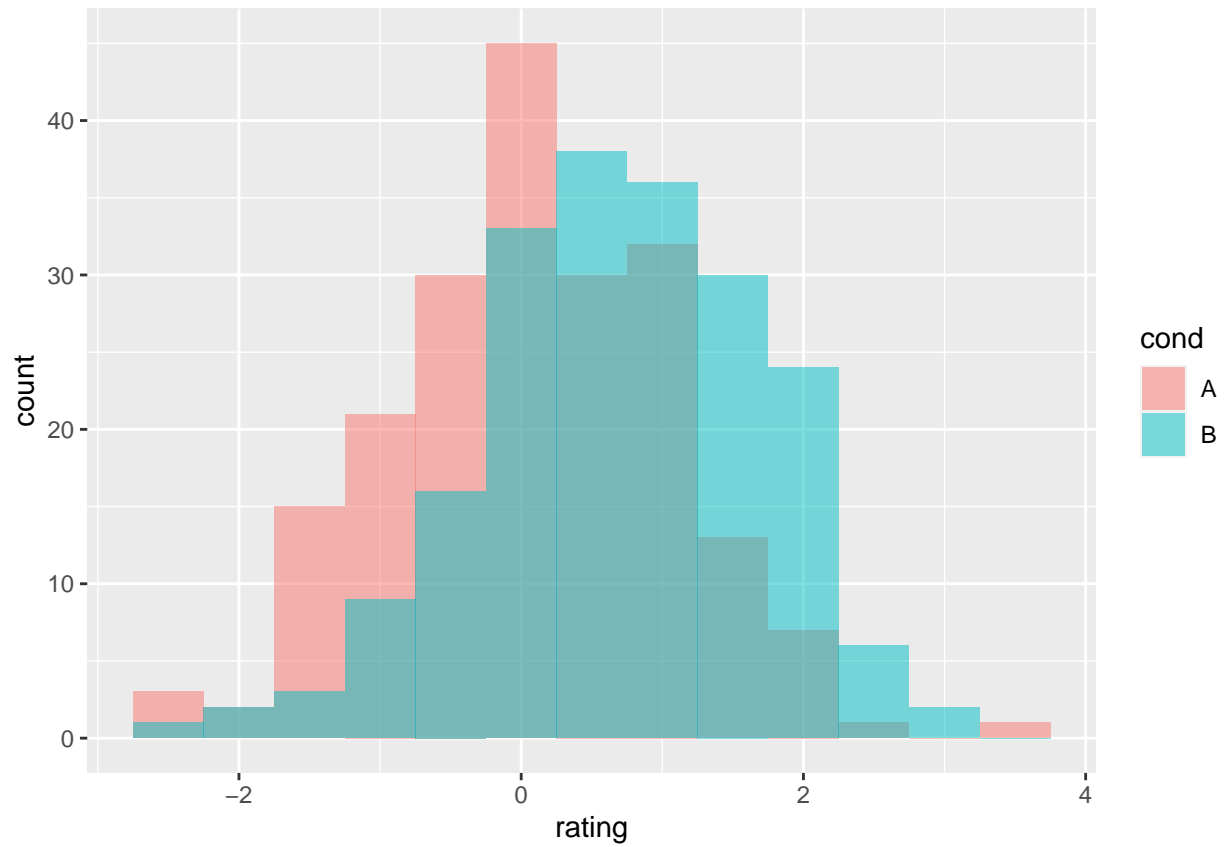
Question 3 - Performing steps to create sample data and observe how each plot function works. Running the following commands given by the Homework for 3a through 3e.

Question 3.a - Creating the sample data.

```
dat <- data.frame(cond = factor(rep(c("A", "B"), each=200)),  
rating = c(rnorm(200), rnorm(200, mean=.8)))
```

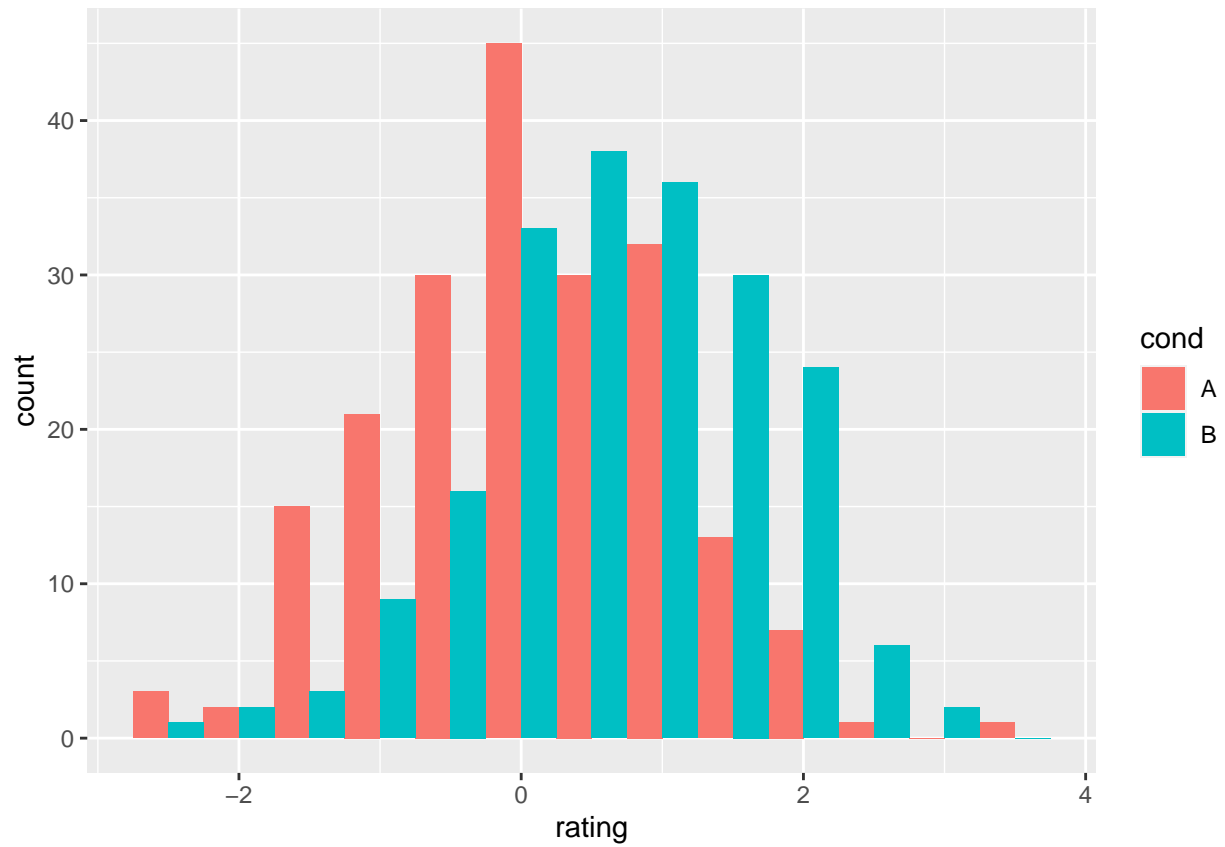
Question 3.b - Overlaid histograms.

```
ggplot(dat, aes(x=rating, fill=cond)) +  
geom_histogram(binwidth=.5, alpha=.5, position="identity")
```



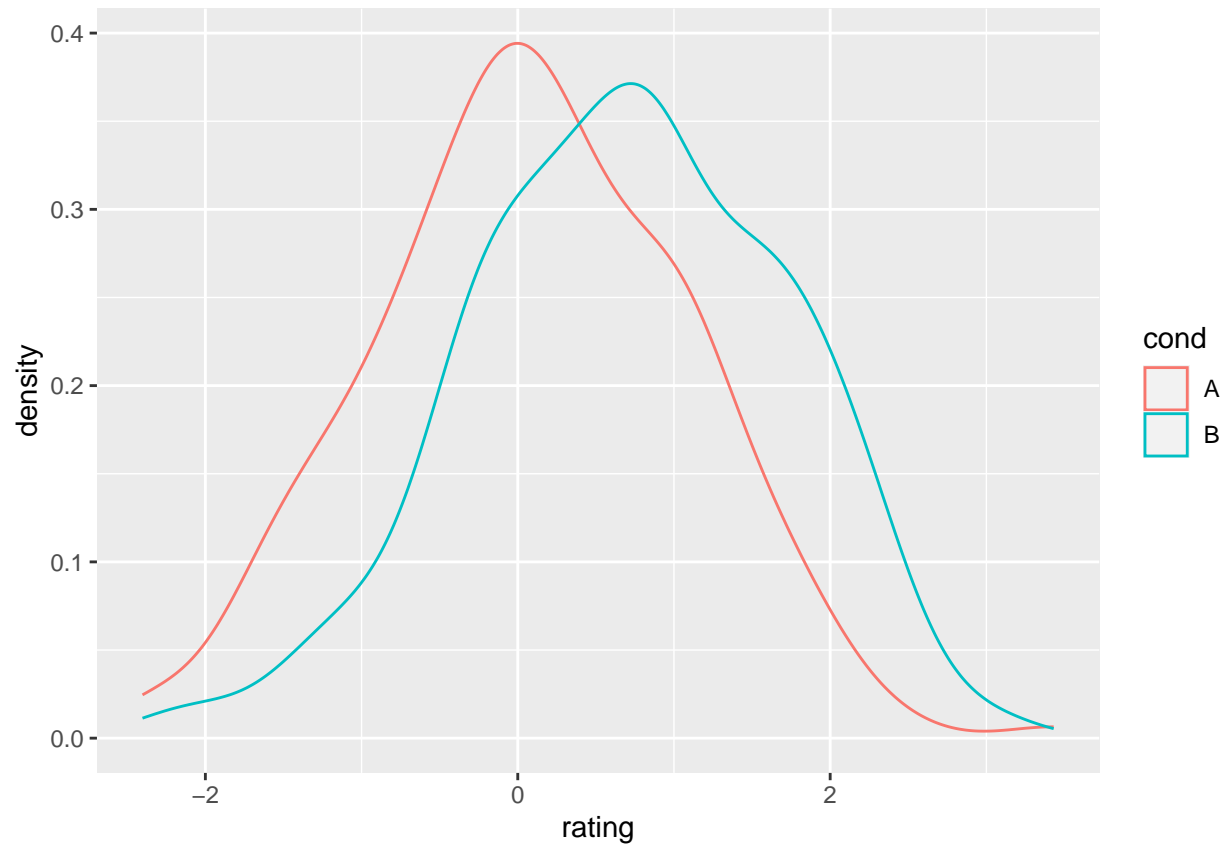
Question 3.c - Interleaved histograms.

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



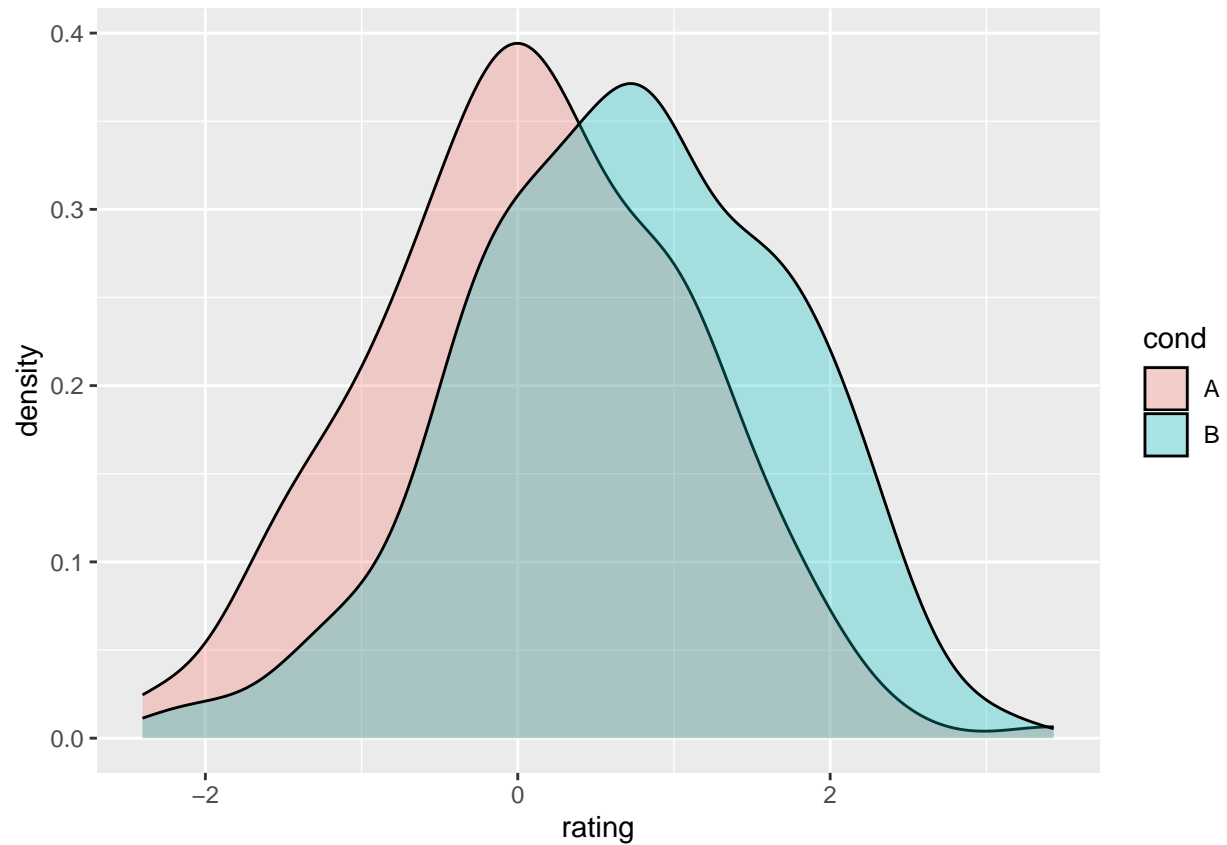
Question 3.d - Density plots.

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



Question 3.e - Density plots with semitransparent fill.

```
ggplot(dat, aes(x=rating, fill=cond)) + geom_density(alpha=.3)
```

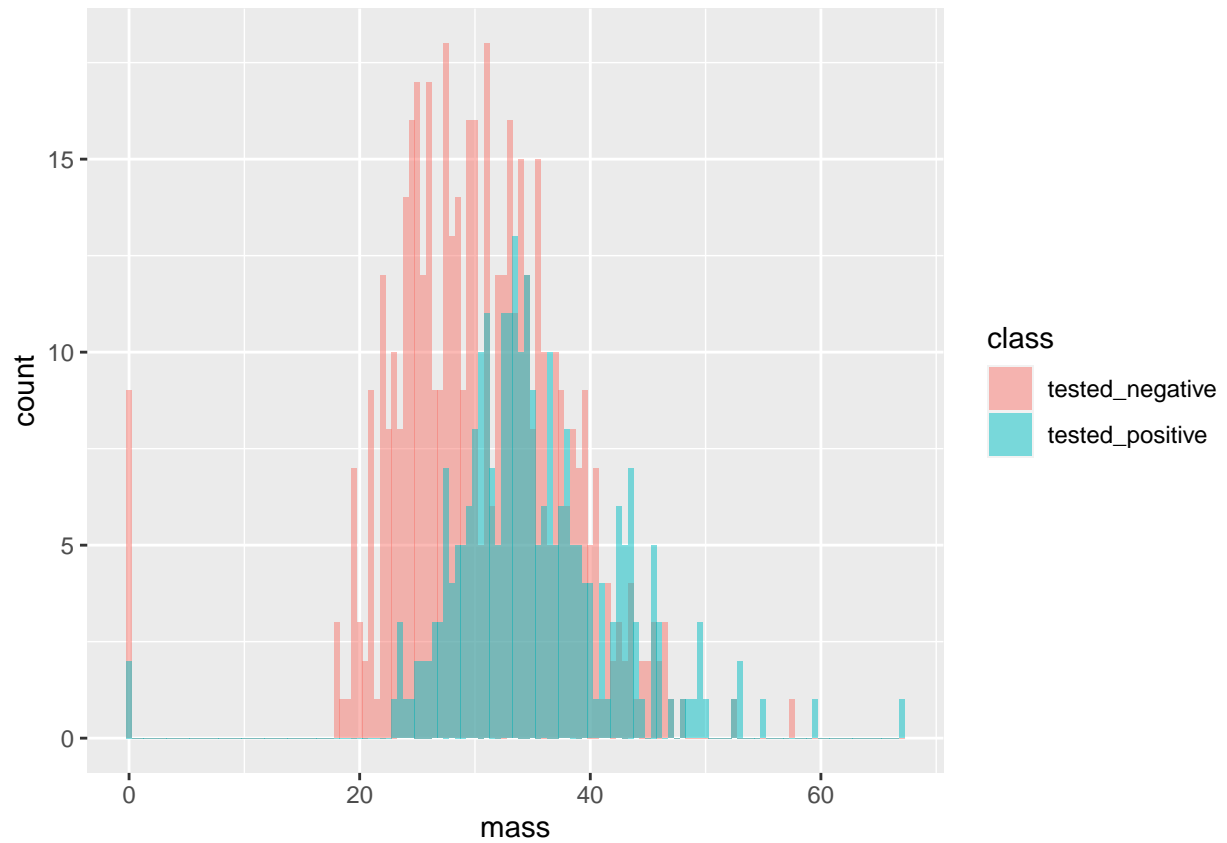


Question 3.f - Reading 'diabetes\_train.csv' into a variable called 'diabetes' and applying the same functions 3.b through 3.e. I first downloaded this data then put it in the 'data' folder.

```
filename2 <- 'data/diabetes_train.csv'
diabetes <- read.csv(filename2, header = TRUE, sep = ',')
```

Overlaid histograms.

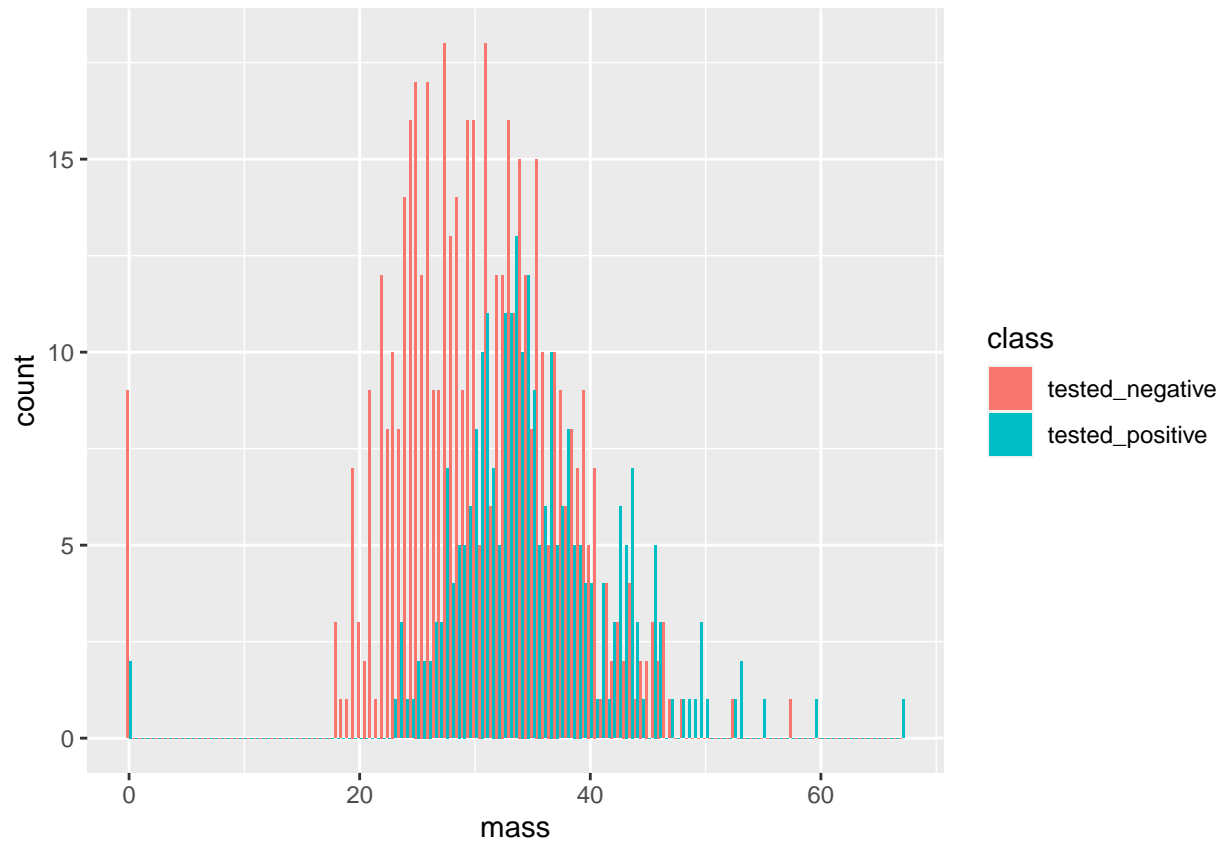
```
ggplot(diabetes, aes(x=mass, fill=class)) +
  geom_histogram(binwidth=.5, alpha=.5, position="identity")
```



Interleaved histograms.

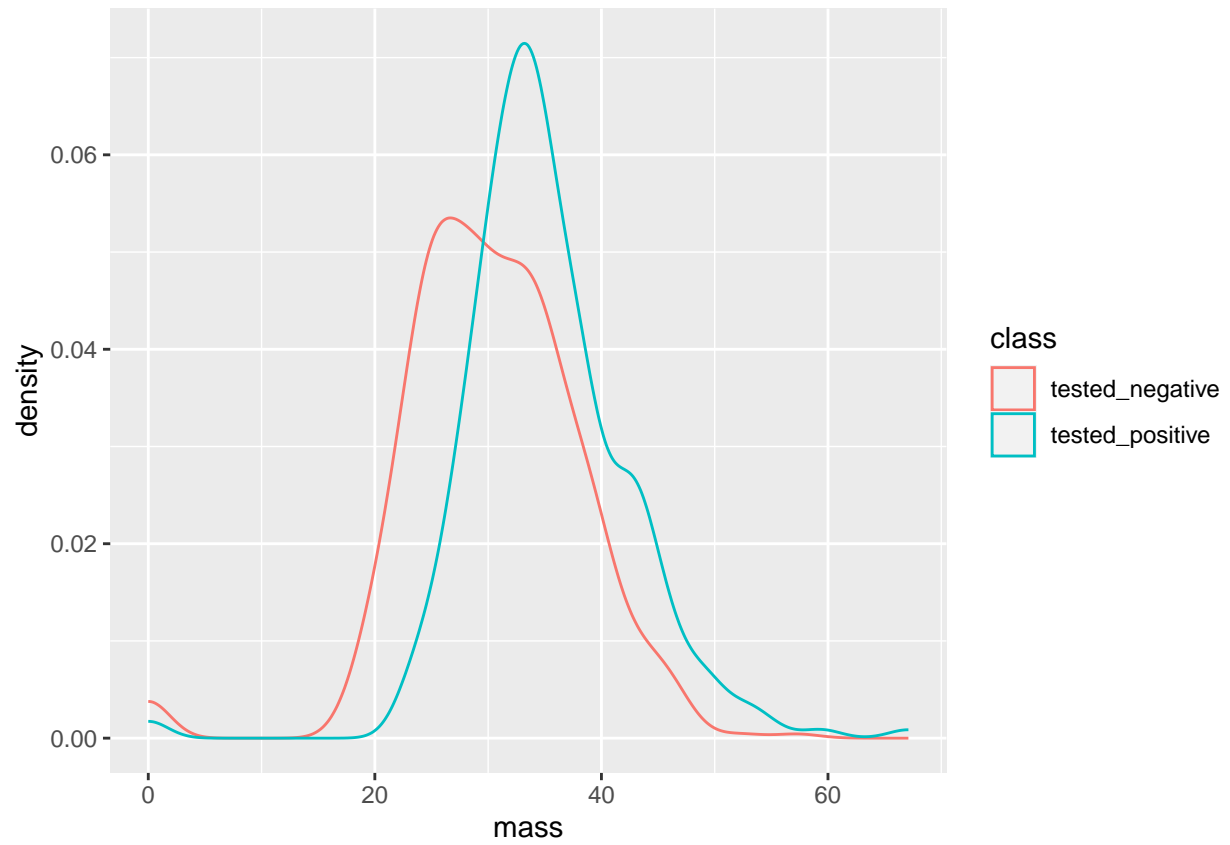
```
ggplot(diabetes, aes(x=mass, fill=class)) + geom_histogram(binwidth=.5, position="dodge")
```





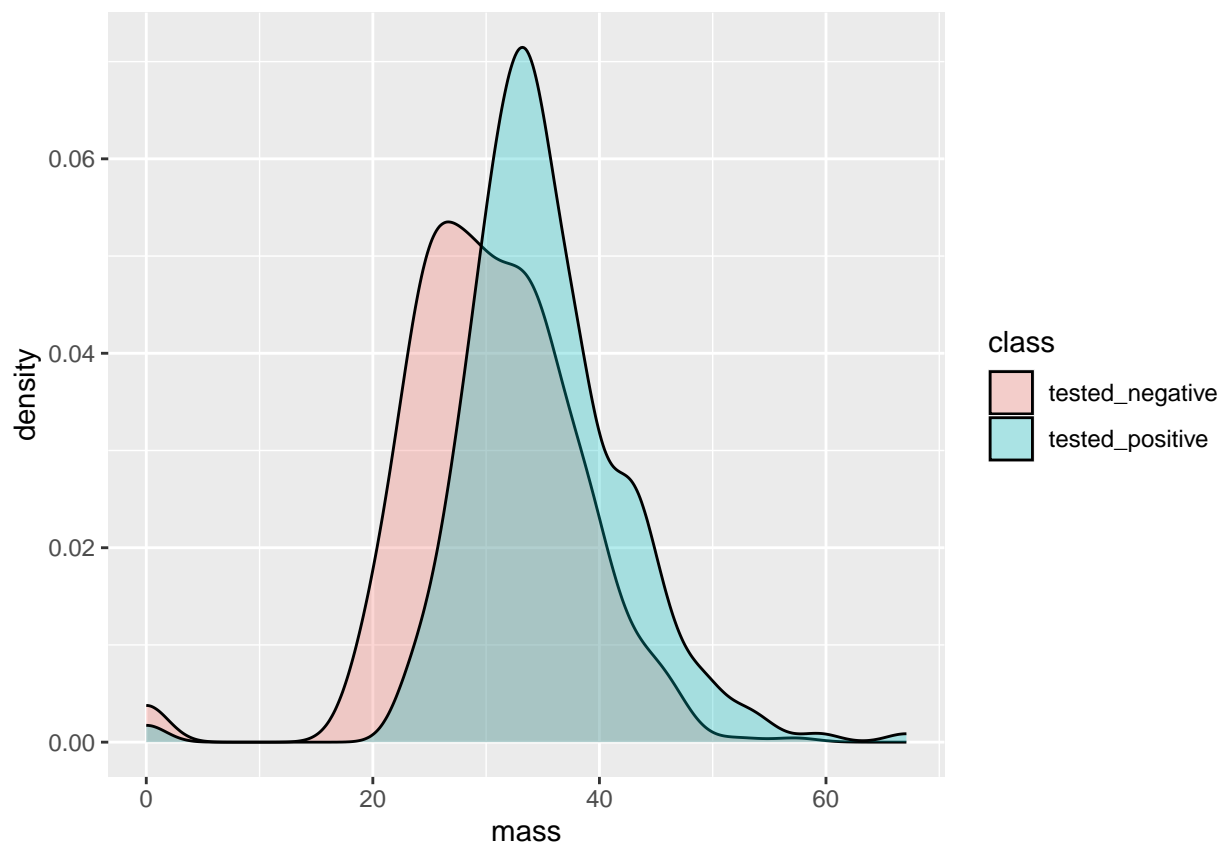
Density plots.

```
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)
```



Question 4 - Read the titanic.csv file to a variable named 'passengers' then perform the following steps and explain the operation.

```
filename3 <- 'data/titanic.csv'
passengers <- read.csv(filename3, header = TRUE, sep = ',')
```

Question 4.a - Shows the summaries of each column in 'passengers' after dropping na values.

```
passengers %>% drop_na() %>% summary()
```

```
##      X      PassengerId      Survived      Pclass
##  Min.   : 0.0   Min.   : 1.0   Min.   :0.0000   Length:714
##  1st Qu.:221.2  1st Qu.:222.2  1st Qu.:0.0000   Class :character
##  Median :444.0  Median :445.0  Median :0.0000   Mode  :character
##  Mean   :447.6  Mean   :448.6  Mean   :0.4062
##  3rd Qu.:676.8  3rd Qu.:677.8  3rd Qu.:1.0000
##  Max.   :890.0  Max.   :891.0  Max.   :1.0000
##      Name      Sex      Age      SibSp
##  Length:714    Length:714    Min.   : 0.42   Min.   :0.0000
##  Class :character  Class :character  1st Qu.:20.12   1st Qu.:0.0000
##  Mode  :character  Mode  :character  Median :28.00   Median :0.0000
##                                     Mean   :29.70   Mean   :0.5126
##                                     3rd Qu.:38.00   3rd Qu.:1.0000
##                                     Max.   :80.00   Max.   :5.0000
##      Parch      Ticket      Fare      Cabin
```

```
## Min. :0.0000 Length:714 Min. : 0.00 Length:714
## 1st Qu.:0.0000 Class :character 1st Qu.: 8.05 Class :character
## Median :0.0000 Mode :character Median : 15.74 Mode :character
## Mean :0.4314 Mean : 34.69
## 3rd Qu.:1.0000 3rd Qu.: 33.38
## Max. :6.0000 Max. :512.33
## Embarked
## Length:714
## Class :character
## Mode :character
##
##
##
```

Question 4.b - Shows the rows with a 'Sex' value of 'male'.

```
passengers %>% filter(Sex == "male") %>% head(10)
```

```
##      X PassengerId Survived Pclass      Name Sex Age SibSp
## 1  0           1         0      3 Braund, Mr. Owen Harris male  22    1
## 2  4           5         0      3 Allen, Mr. William Henry male  35    0
## 3  5           6         0      3 Moran, Mr. James male  NA    0
## 4  6           7         0      1 McCarthy, Mr. Timothy J male  54    0
## 5  7           8         0      3 Palsson, Master. Gosta Leonard male  2    3
## 6 12          13         0      3 Saundercock, Mr. William Henry male  20    0
## 7 13          14         0      3 Andersson, Mr. Anders Johan male  39    1
## 8 16          17         0      3 Rice, Master. Eugene male  2    4
## 9 17          18         1      2 Williams, Mr. Charles Eugene male  NA    0
## 10 20         21         0      2 Fynney, Mr. Joseph J male  35    0
##      Parch Ticket   Fare Cabin Embarked
## 1      0 A/5 21171  7.2500      S
## 2      0 373450  8.0500      S
## 3      0 330877  8.4583      Q
## 4      0 17463 51.8625   E46      S
## 5      1 349909 21.0750      S
## 6      0 A/5. 2151  8.0500      S
## 7      5 347082 31.2750      S
## 8      1 382652 29.1250      Q
## 9      0 244373 13.0000      S
## 10     0 239865 26.0000      S
```

Question 4.c - Lists the dataframe rows in descending order of 'Fare' (most to least expensive.)

```
passengers %>% arrange(desc(Fare)) %>% head(10)
```

```
##      X PassengerId Survived Pclass      Name Sex
## 1 258          259         1      1 Ward, Miss. Anna female
## 2 679          680         1      1 Cardeza, Mr. Thomas Drake Martinez male
## 3 737          738         1      1 Lesurer, Mr. Gustave J male
## 4 27           28         0      1 Fortune, Mr. Charles Alexander male
## 5 88           89         1      1 Fortune, Miss. Mabel Helen female
## 6 341          342         1      1 Fortune, Miss. Alice Elizabeth female
```

```
## 7 438      439      0      1      Fortune, Mr. Mark  male
## 8 311      312      1      1      Ryerson, Miss. Emily Borie female
## 9 742      743      1      1 Ryerson, Miss. Susan Parker "Suzette" female
## 10 118     119      0      1      Baxter, Mr. Quigg Edmond  male
##   Age SibSp Parch   Ticket   Fare      Cabin Embarked
## 1  35     0     0 PC 17755 512.3292      C
## 2  36     0     1 PC 17755 512.3292    B51 B53 B55      C
## 3  35     0     0 PC 17755 512.3292    B101      C
## 4  19     3     2  19950 263.0000    C23 C25 C27      S
## 5  23     3     2  19950 263.0000    C23 C25 C27      S
## 6  24     3     2  19950 263.0000    C23 C25 C27      S
## 7  64     1     4  19950 263.0000    C23 C25 C27      S
## 8  18     2     2 PC 17608 262.3750 B57 B59 B63 B66      C
## 9  21     2     2 PC 17608 262.3750 B57 B59 B63 B66      C
## 10 24     0     1 PC 17558 247.5208    B58 B60      C
```

Question 4.d - Creates a new column called 'FamSize' (family size) that is equal to Parch (number of parents/children) + SibSp (number of siblings), then displays rows with the new column.

```
passengers %>% mutate(FamSize = Parch + SibSp) %>% head(10)
```

```
##   X PassengerId Survived Pclass
## 1 0           1         0      3
## 2 1           2         1      1
## 3 2           3         1      3
## 4 3           4         1      1
## 5 4           5         0      3
## 6 5           6         0      3
## 7 6           7         0      1
## 8 7           8         0      3
## 9 8           9         1      3
## 10 9          10         1      2
##                                     Name   Sex Age SibSp Parch
## 1                               Braund, Mr. Owen Harris  male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                               Heikkinen, Miss. Laina female  26     0     0
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                               Allen, Mr. William Henry  male  35     0     0
## 6                               Moran, Mr. James         male  NA     0     0
## 7                               McCarthy, Mr. Timothy J  male  54     0     0
## 8                               Palsson, Master. Gosta Leonard  male   2     3     1
## 9 Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female  27     0     2
## 10 Nasser, Mrs. Nicholas (Adele Achem) female  14     1     0
##   Ticket   Fare Cabin Embarked FamSize
## 1  A/5 21171  7.2500      S      1
## 2  PC 17599 71.2833    C85      1
## 3 STON/O2. 3101282  7.9250      S      0
## 4  113803 53.1000   C123      1
## 5  373450  8.0500      S      0
## 6  330877  8.4583      Q      0
## 7  17463 51.8625    E46      0
## 8  349909 21.0750      S      4
```

```
## 9          347742 11.1333          S          2
## 10         237736 30.0708          C          1
```

Question 4.e - Groups the rows by 'Sex' then shows the mean 'fare' and number of survivors for each 'Sex.'

```
passengers %>% group_by(Sex) %>% summarise(meanFare = mean(Fare), numSurv = sum(Survived))
```

```
## # A tibble: 2 x 3
##   Sex      meanFare numSurv
##   <chr>      <dbl>   <int>
## 1 female    44.5     233
## 2 male     25.5     109
```

Question 5- Using the 'quantile' function to calculate the 10th, 30th, 50th and 60th percentiles of skin attribute of diabetes data.

```
quantile(diabetes$skin, probs=c(0.1, 0.3, 0.5, 0.6))
```

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