CSC487-HW1

Corey Malone

2025-01-29

1. Work with "Su_raw_matrix.txt"

```
library(ggplot2)

# (a) Read data using read.delim
#    'header=TRUE' means we assume the first row is column names.
su <- read.delim("Su_raw_matrix.txt", header = TRUE)

# Let's quickly check the structure (optional)
str(su)
```

```
## 'data.frame':    12626 obs. of  8 variables:
##  $ Brain_1.CEL      : num  120.25 583.6 35.85 17.6 0.15 ...
##  $ Brain_2.CEL      : num  255 885.4 40.5 19.9 26.4 ...
##  $ Fetal_brain_1.CEL: num  3.5 253.7 47.2 11.1 78 ...
##  $ Fetal_brain_2.CEL: num  31 293.4 33 23.1 36 ...
##  $ Fetal_liver_1.CEL: num  6.5 201.2 86.3 38.8 89.5 ...
##  $ Fetal_liver_2.CEL: num  -8.25 433.75 119.25 94.6 34 ...
##  $ Liver_1.CEL      : num  19.1 134.2 37.1 452.1 22.8 ...
##  $ Liver_2.CEL      : num  73 251.2 72.1 662.5 100 ...
```

```
# (b) Compute mean and standard deviation of Liver_2.CEL column
mean_Liver2 <- mean(su$Liver_2.CEL, na.rm = TRUE)  # na.rm=TRUE to ignore any NA values
sd_Liver2   <- sd(su$Liver_2.CEL, na.rm = TRUE)

# Print results
cat("Mean of Liver_2.CEL:", mean_Liver2, "\n")
```

```
## Mean of Liver_2.CEL: 241.8246
```

```
cat("SD of Liver_2.CEL:", sd_Liver2, "\n")
```

```
## SD of Liver_2.CEL: 1133.352
```

```
# (c) Get average (colMeans) and total (colSums) values for each column
column_means <- colMeans(su, na.rm = TRUE)
column_sums  <- colSums(su, na.rm = TRUE)

cat("\nColumn means:\n")
```

```
##
## Column means:
```

```
print(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
```

```
cat("\nColumn sums:\n")
```

```
##
## Column sums:
```

```
print(column_sums)
```

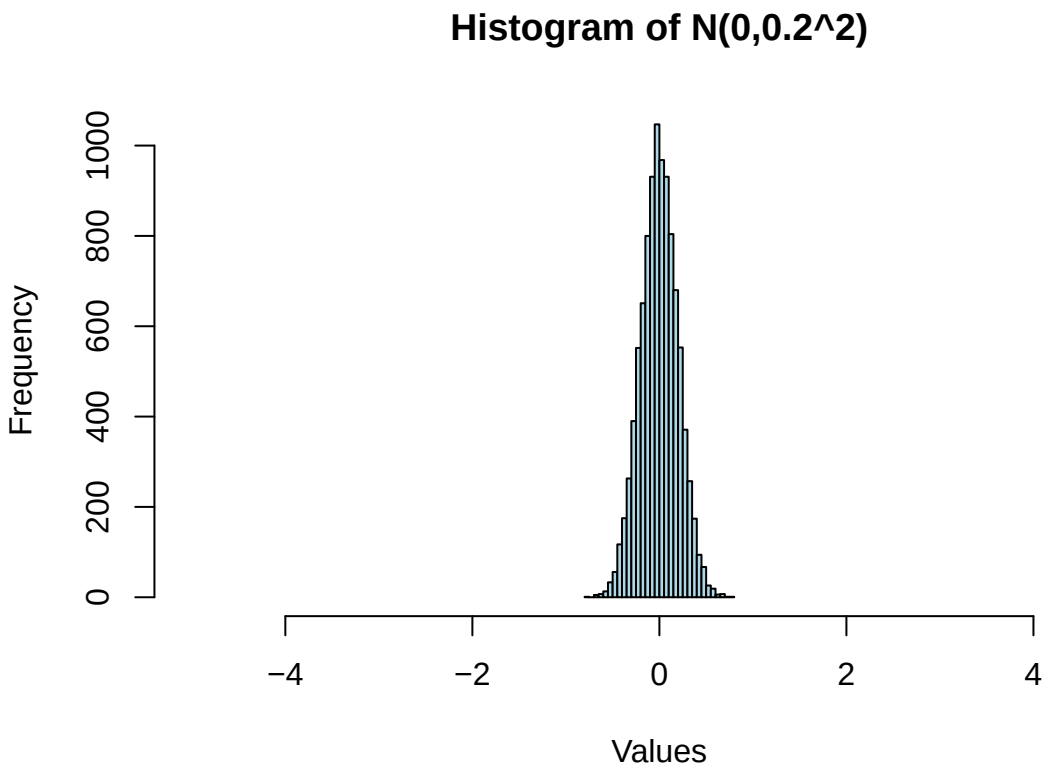
```
##       Brain_1.CEL       Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL
##           2588031           3978357           2504290           3379413
## Fetal_liver_1.CEL Fetal_liver_2.CEL       Liver_1.CEL       Liver_2.CEL
##           2649846           5039645           2030966           3053278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

2. Generate random numbers from normal distributions and plot histograms

We'll generate 10,000 random values from N(0, 0.2^2) and N(0, 0.5^2) Then plot histograms and comment on their differences.

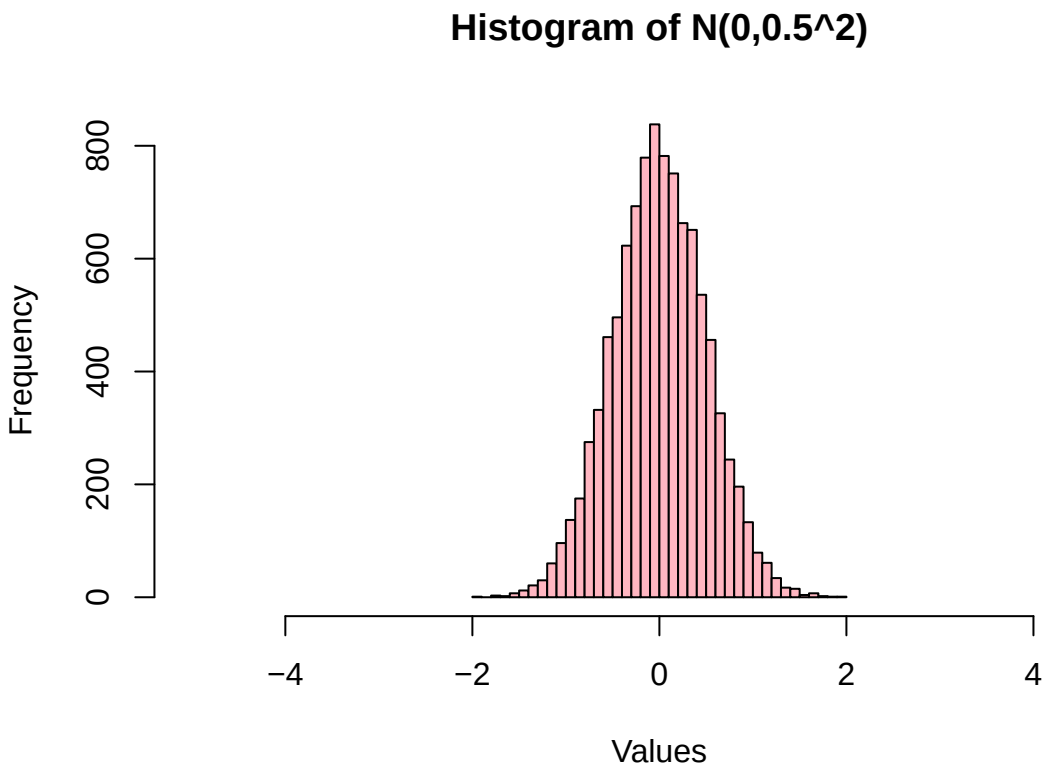
```
# (a) mean=0, sigma=0.2
set.seed(123)  # Optional, for reproducible results
x1 <- rnorm(10000, mean = 0, sd = 0.2)

hist(x1,
     main = "Histogram of N(0,0.2^2)",
     xlab = "Values",
     xlim = c(-5,5),   # so we can compare easily
     breaks = 50,
     col = "lightblue")
```




```r
# (b) mean=0, sigma=0.5
set.seed(123)
x2 <- rnorm(10000, mean = 0, sd = 0.5)

hist(x2,
     main = "Histogram of N(0,0.5^2)",
     xlab = "Values",
     xlim = c(-5,5),
     breaks = 50,
     col = "lightpink")
```



Comment on differences: The distribution with sigma=0.2 is narrower (less spread), while sigma=0.5 is wider (more spread). Both have mean 0.

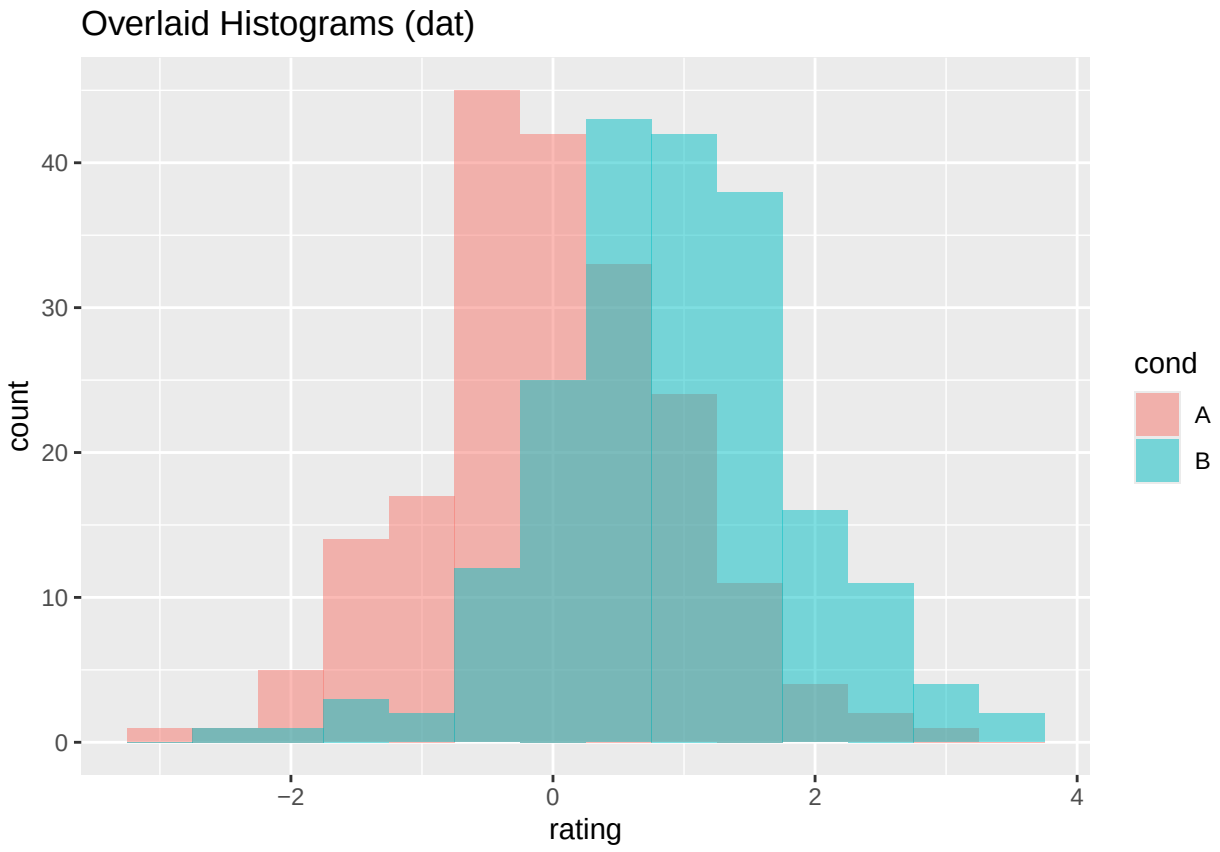
3. Use ggplot2 with "dat" and then "diabetes"

First, create a sample dataframe 'dat'

```
dat <- data.frame(
  cond = factor(rep(c("A", "B"), each = 200)),
  rating = c(rnorm(200), rnorm(200, mean = 0.8))
)
```

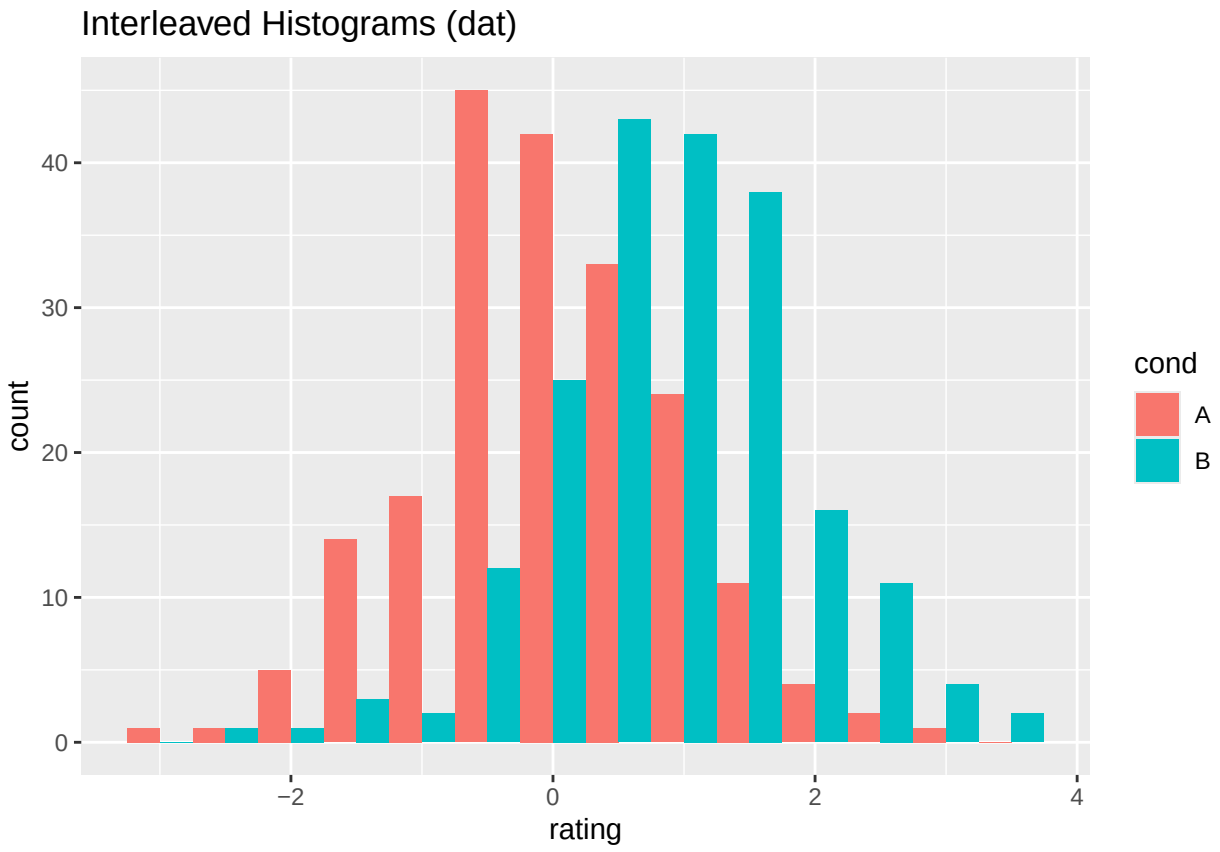
(b) Overlaid histograms

```
ggplot(dat, aes(x = rating, fill = cond)) +
  geom_histogram(binwidth = 0.5, alpha = 0.5, position = "identity") +
  ggtitle("Overlaid Histograms (dat)")
```



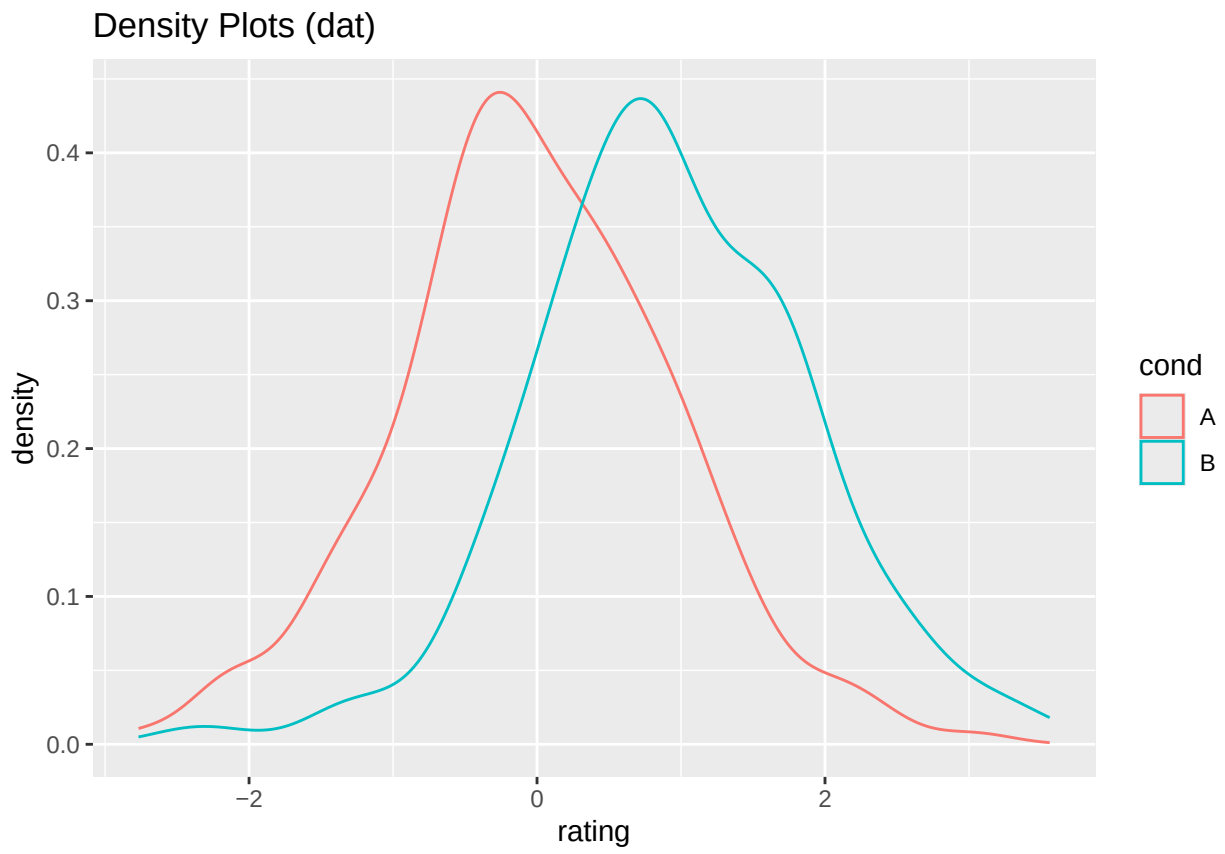
(c) Interleaved histograms

```
ggplot(dat, aes(x = rating, fill = cond)) +
  geom_histogram(binwidth = 0.5, position = "dodge") +
  ggtitle("Interleaved Histograms (dat)")
```



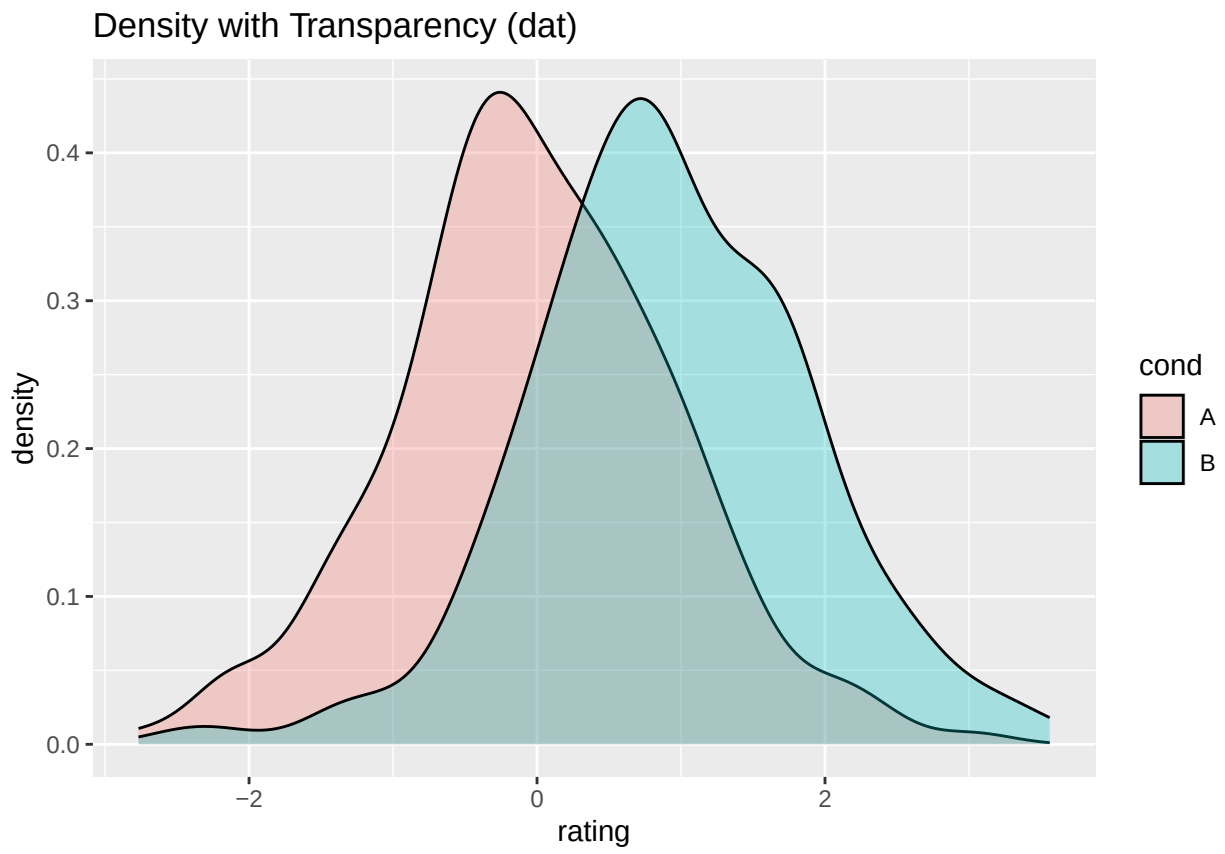
(d) Density plots

```
ggplot(dat, aes(x = rating, colour = cond)) +
  geom_density() +
  ggtitle("Density Plots (dat)")
```



(e) Density plots with semitransparent fill

```
ggplot(dat, aes(x = rating, fill = cond)) +
  geom_density(alpha = 0.3) +
  ggtitle("Density with Transparency (dat)")
```



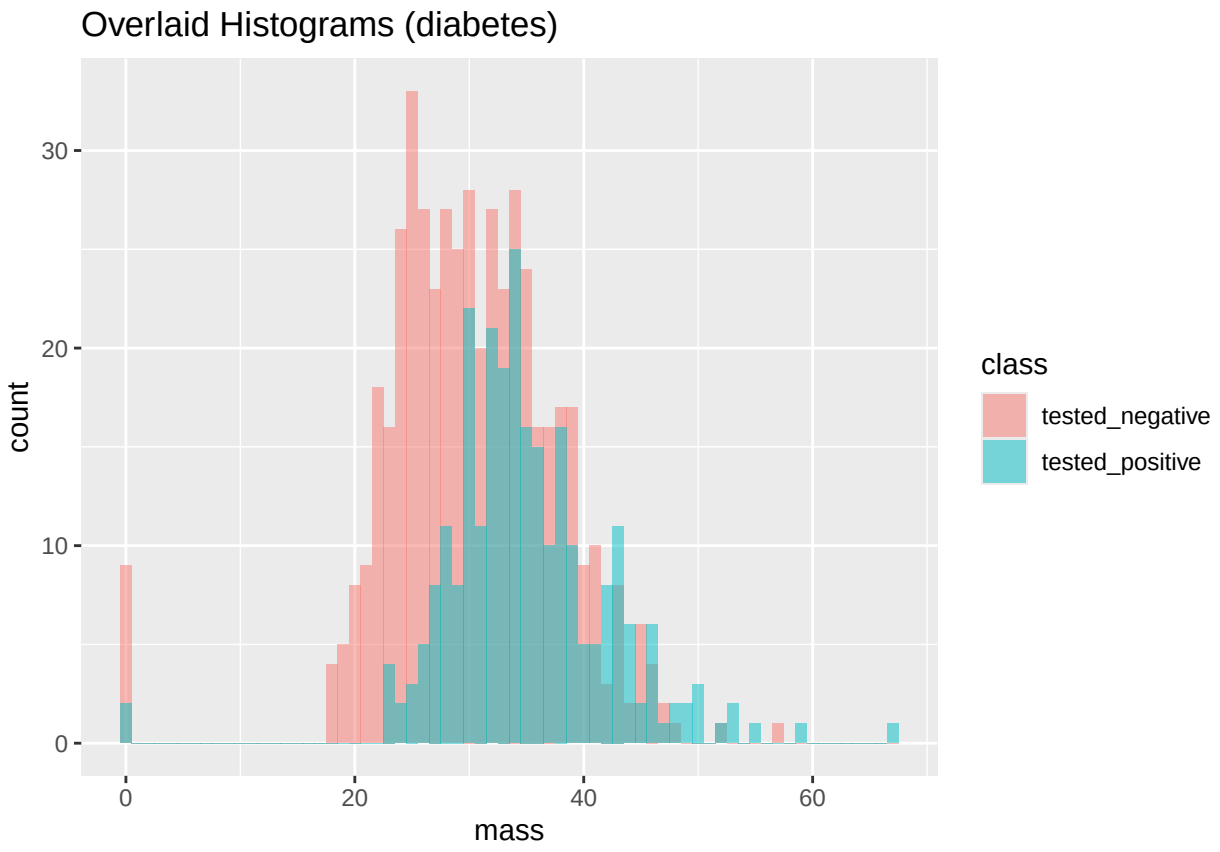
(f) Read "diabetes_train.csv" into 'diabetes' and apply same functions for 'mass' attribute, grouping by 'class'

```
diabetes <- read.csv("diabetes_train.csv", header = TRUE)
str(diabetes)  # Check structure to see the column names
```

```
## 'data.frame':    758 obs. of  9 variables:
##  $ preg : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ plas : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ pres : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ skin : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ insu : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ mass : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ pedi : num  0.627 0.351 0.672 0.167 2.288 ...
##  $ age  : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ class: chr  "tested_positive" "tested_negative" "tested_positive" "tested_negative" ...
```

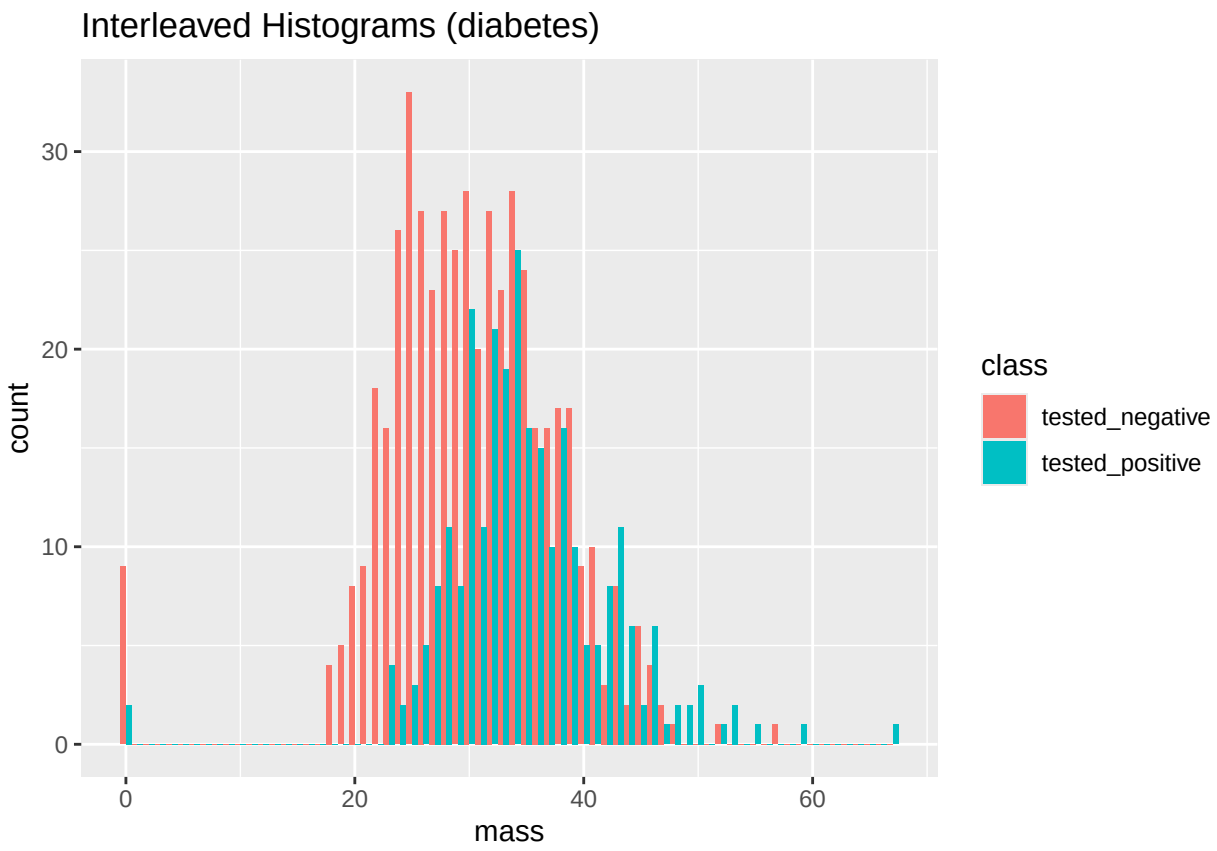
Overlaid histograms for 'mass' by 'class'

```
ggplot(diabetes, aes(x = mass, fill = class)) +
  geom_histogram(binwidth = 1, alpha = 0.5, position = "identity") +
  ggtitle("Overlaid Histograms (diabetes)")
```



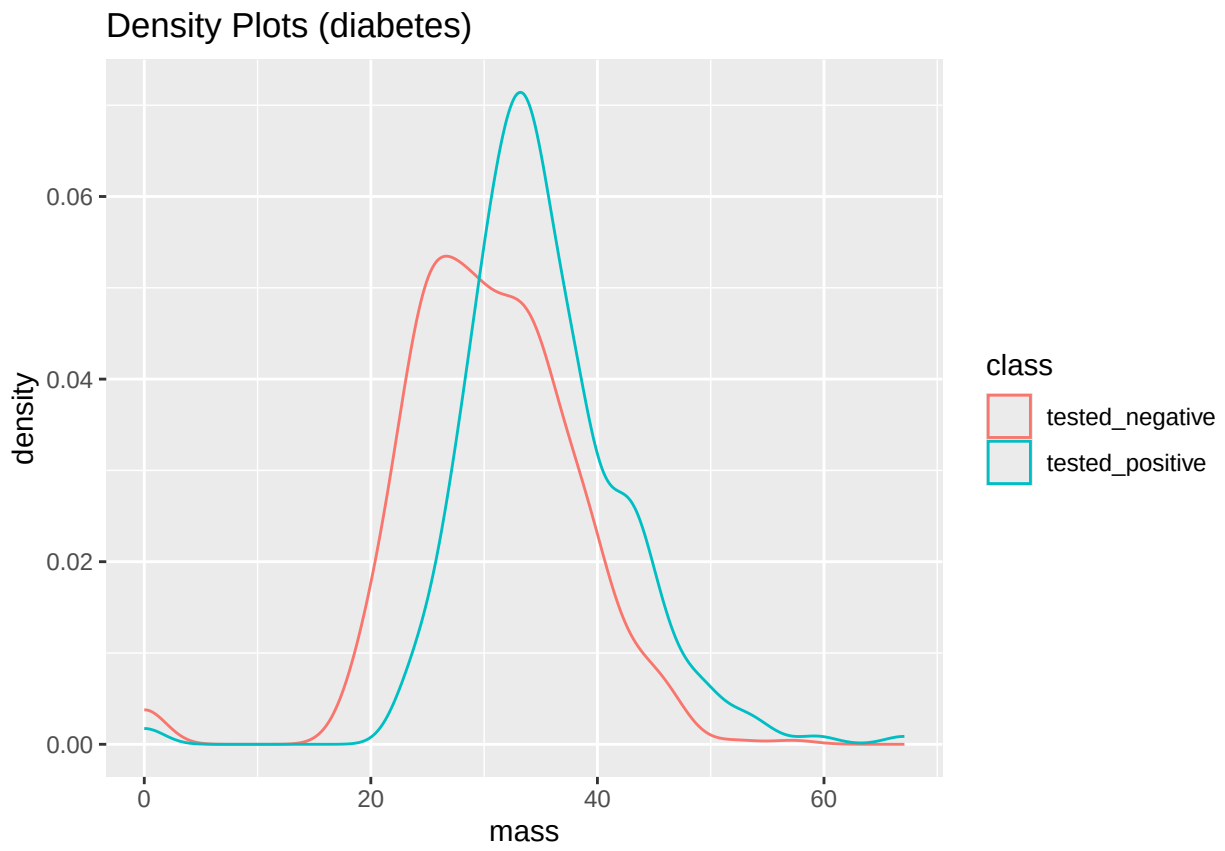
Interleaved histograms

```
ggplot(diabetes, aes(x = mass, fill = class)) +
  geom_histogram(binwidth = 1, position = "dodge") +
  ggtitle("Interleaved Histograms (diabetes)")
```



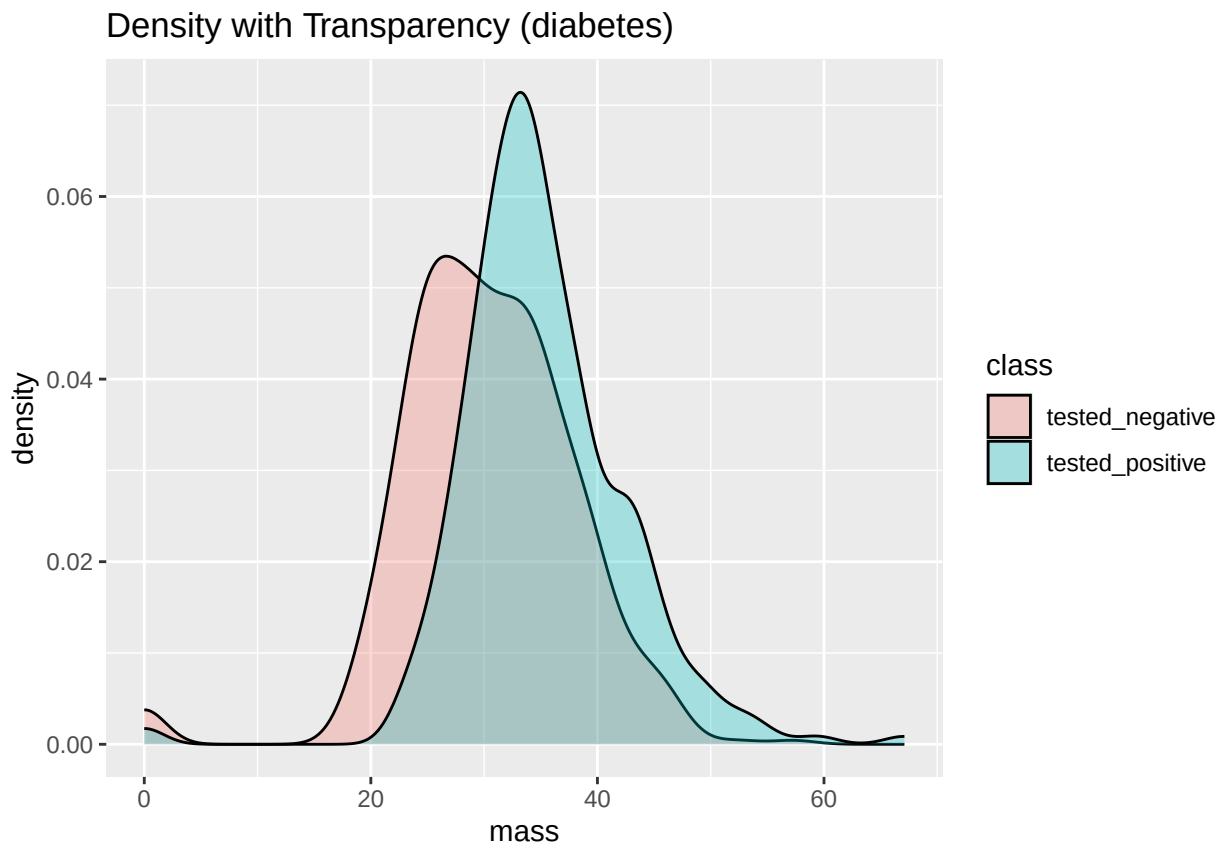
Density plots

```
ggplot(diabetes, aes(x = mass, colour = class)) +
  geom_density() +
  ggtitle("Density Plots (diabetes)")
```



Density plots with semitransparent fill

```
ggplot(diabetes, aes(x = mass, fill = class)) +
  geom_density(alpha = 0.3) +
  ggtitle("Density with Transparency (diabetes)")
```



4. Read "titanic.csv" and do base R operations

```
passengers <- read.csv("titanic.csv", header = TRUE)
```

(a) Drop NA rows and get summary

```
passengers_noNA <- na.omit(passengers)  # Removes rows with any NA

summary(passengers_noNA)
```

```
##        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
##
##
##
```

(b) Filter (Sex == "male") using base R subset

```r
passengers_male <- subset(passengers, Sex == "male")
# Print first rows (optional)
head(passengers_male)
```

```
##     X PassengerId Survived Pclass                            Name  Sex Age SibSp
## 1   0           1        0      3         Braund, Mr. Owen Harris male  22     1
## 5   4           5        0      3        Allen, Mr. William Henry male  35     0
## 6   5           6        0      3                Moran, Mr. James male  NA     0
## 7   6           7        0      1         McCarthy, Mr. Timothy J male  54     0
## 8   7           8        0      3 Palsson, Master. Gosta Leonard male   2     3
## 13 12          13        0      3 Saundercock, Mr. William Henry male  20     0
##    Parch    Ticket    Fare Cabin Embarked
## 1      0 A/5 21171  7.2500              S
## 5      0    373450  8.0500              S
## 6      0    330877  8.4583              Q
## 7      0     17463 51.8625   E46        S
## 8      1    349909 21.0750              S
## 13     0 A/5. 2151  8.0500              S
```

(c) Arrange in descending order by Fare

```r
passengers_descFare <- passengers[order(-passengers$Fare), ]
head(passengers_descFare)
```

```
##       X PassengerId Survived Pclass                                  Name    Sex
## 259 258         259        1      1                    Ward, Miss. Anna female
## 680 679         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male
```

```
## 738 737         738        1      1            Lesurer, Mr. Gustave J   male
## 28   27          28        0      1     Fortune, Mr. Charles Alexander   male
## 89   88          89        1      1         Fortune, Miss. Mabel Helen female
## 342 341         342        1      1     Fortune, Miss. Alice Elizabeth female
##     Age SibSp Parch   Ticket     Fare       Cabin Embarked
## 259  35     0     0 PC 17755 512.3292                   C
## 680  36     0     1 PC 17755 512.3292 B51 B53 B55       C
## 738  35     0     0 PC 17755 512.3292        B101       C
## 28   19     3     2    19950 263.0000 C23 C25 C27       S
## 89   23     3     2    19950 263.0000 C23 C25 C27       S
## 342  24     3     2    19950 263.0000 C23 C25 C27       S
```

(d) Mutate: Add FamSize = Parch + SibSp

```
passengers$FamSize <- passengers$Parch + passengers$SibSp
head(passengers)
```

```
##   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
##                                                 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
##             Ticket    Fare Cabin Embarked FamSize
## 1        A/5 21171  7.2500              S       1
## 2         PC 17599 71.2833   C85        C       1
## 3 STON/O2. 3101282  7.9250              S       0
## 4           113803 53.1000  C123        S       1
## 5           373450  8.0500              S       0
## 6           330877  8.4583              Q       0
```

(e) Group by Sex, summarise mean(Fare) and sum(Survived)

We'll use tapply in base R.

```
meanFare_bySex <- tapply(passengers$Fare, passengers$Sex, mean, na.rm = TRUE)
numSurv_bySex  <- tapply(passengers$Survived, passengers$Sex, sum, na.rm = TRUE)

# Combine results in a small data frame
```

```r
groupedResults <- data.frame(
  Sex           = names(meanFare_bySex),
  meanFare      = as.numeric(meanFare_bySex),
  totalSurvived = as.numeric(numSurv_bySex)
)

groupedResults
```

```
##      Sex meanFare totalSurvived
## 1 female 44.47982           233
## 2   male 25.52389           109
```

5. Quantiles of 'skin' in the diabetes data

We need the 10th, 30th, 50th, 60th percentiles

```r
skin_quantiles <- quantile(diabetes$skin, probs = c(0.1, 0.3, 0.5, 0.6), na.rm = TRUE)
cat("\nQuantiles (10%, 30%, 50%, 60%) of 'skin' in diabetes:\n")
```

```
##
## Quantiles (10%, 30%, 50%, 60%) of 'skin' in diabetes:
```

```r
print(skin_quantiles)
```

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