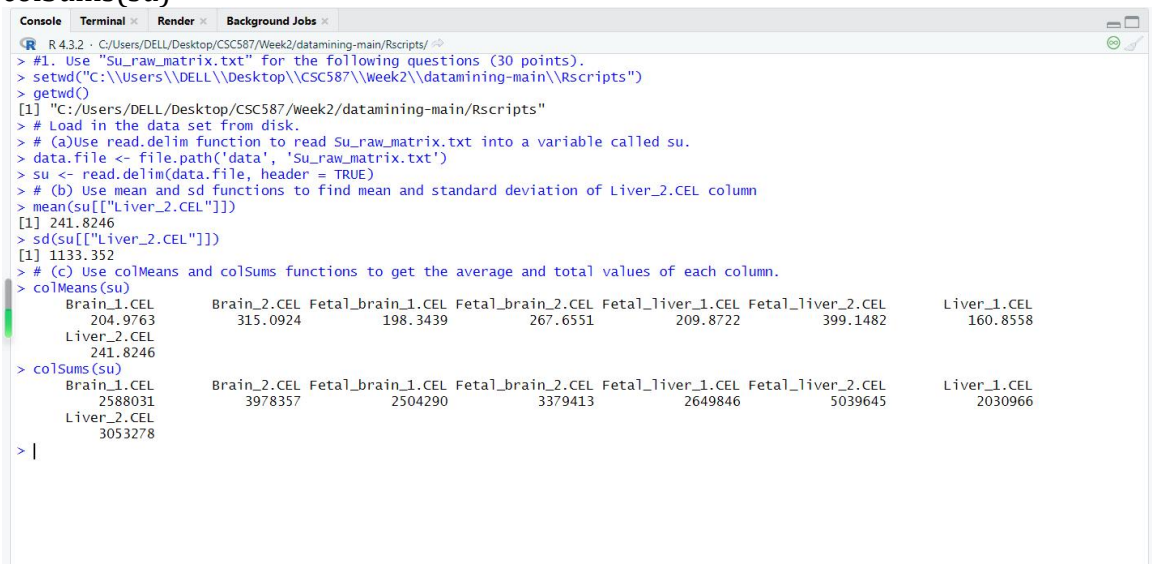
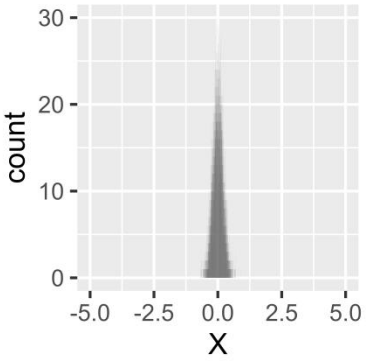
1. Use ”Su_raw_matrix.txt” for the following questions (30 points).
setwd("C:\\Users\\DELL\\Desktop\\CSC587\\Week2\\datamining-main\\Rscripts")
getwd()
(a)Use read.delim function to read Su_raw_matrix.txt into a variable called su.
data.file <- file.path('data', 'Su_raw_matrix.txt')
su <- read.delim(data.file, header = TRUE)
(b) Use mean and sd functions to find mean and standard deviation of Liver_2.CEL column
mean(su[["Liver_2.CEL"]])
sd(su[["Liver_2.CEL"]])
(c) Use colMeans and colSums functions to get the average and total values of each column.
colMeans(su)
colSums(su)



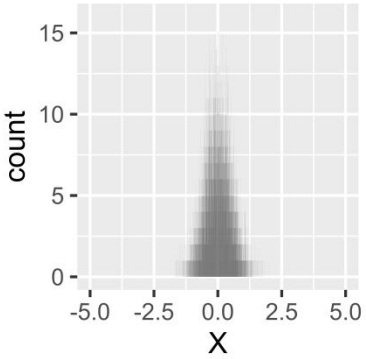
2. Use rnorm(n, mean = 0, sd = 1) function in R to generate 10000 numbers for the following (mean, sigma) pairs and plot histogram for each, meaning you need to change the function parameter accordingly. Then comment on how these histograms are different from each other and state the reason. (20 points)
(a) mean=0, sigma=0.2
sigma1 <- data.frame(X = rnorm(10000, mean = 0, sd = 0.2))
(b) mean=0, sigma=0.5
sigma2 <- data.frame(X = rnorm(10000, mean = 0, sd = 0.5))

Please save your figures as image from RStudio. (Hint: to see the difference in plots you may need to set the xlim parameter in plot function to c(-5,5))
#Start visualizing data using the ggplot2 package.
library('ggplot2')
sigma1ggpot = ggplot(sigma1, aes(x = X)) + geom_histogram(binwidth = 0.001) + xlim(c(-5, 5))

sigma2ggpot = ggplot(sigma2, aes(x = X)) + geom_histogram(binwidth = 0.001) + xlim(c(-5, 5))
ggsave("histogram_sigma1.png", plot = sigma1ggpot, width = 2, height = 2, dpi = 5000)
ggsave("histogram_sigma2.png", plot = sigma2ggpot, width = 2, height = 2, dpi = 5000)



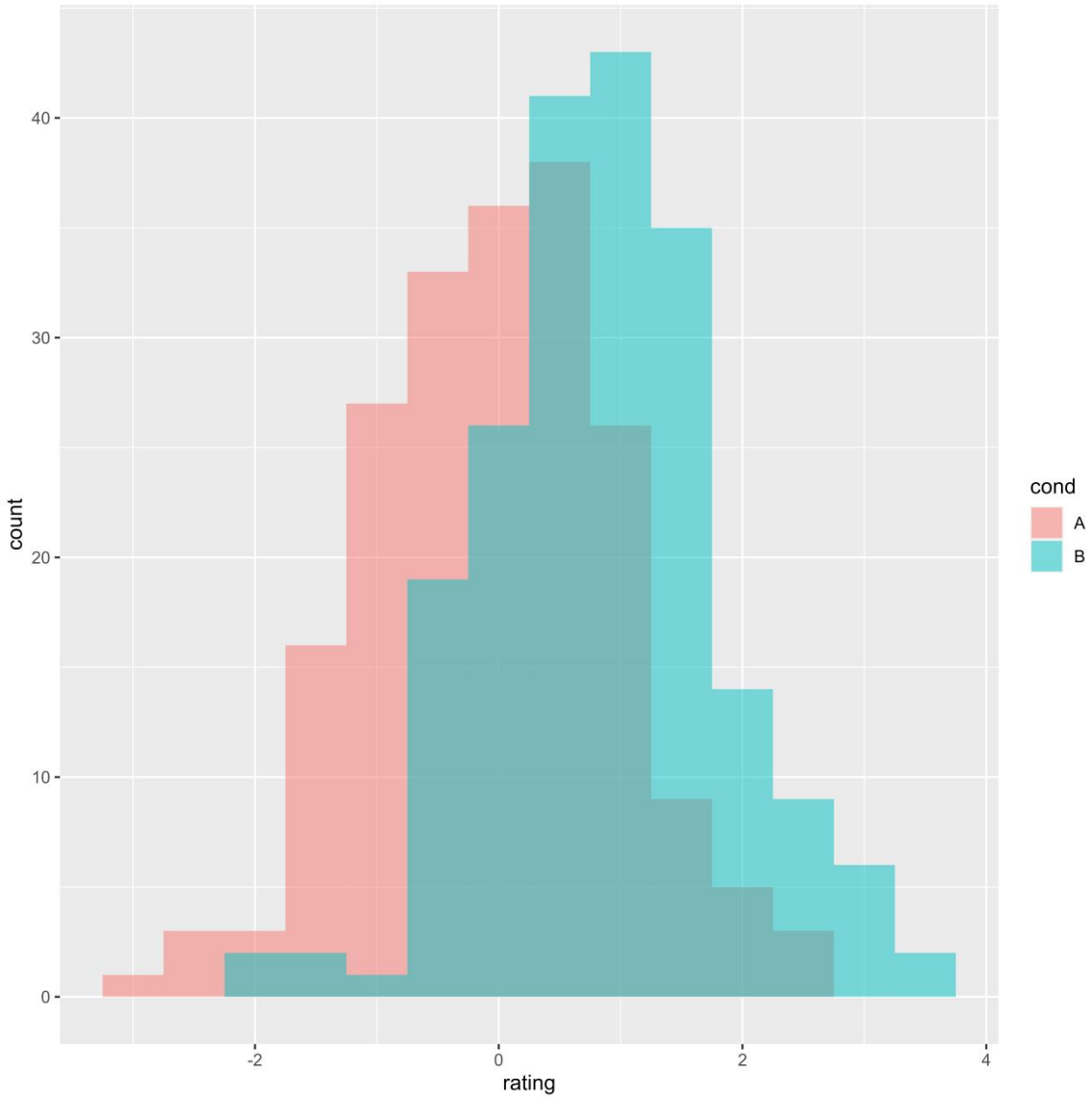
(a) mean=0, sigma=0.2



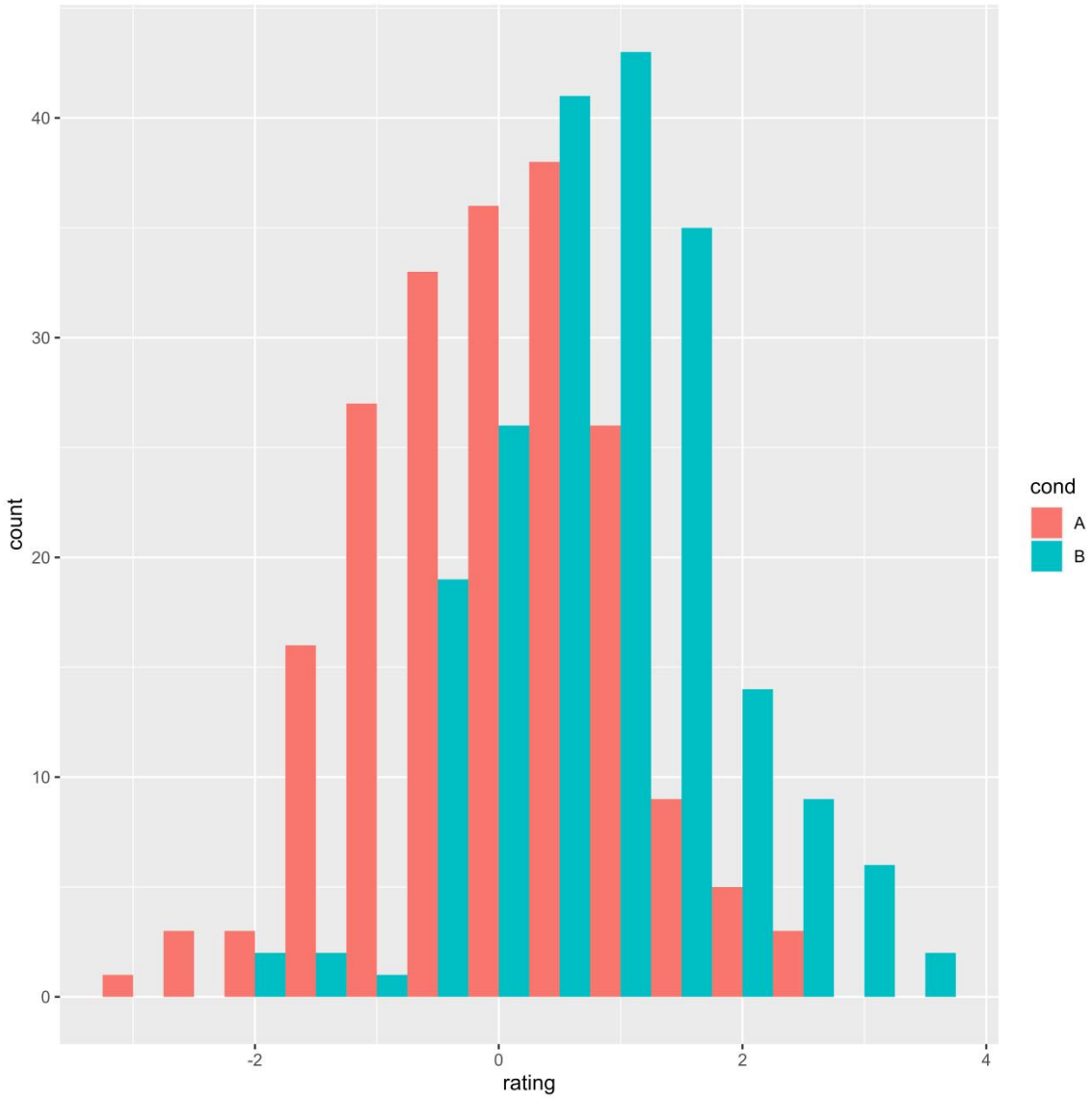
(b) mean=0, sigma=0.5
The sigma 0.5 is lower/shorter and wider. The reason for these differences is that the standard deviation (sigma) controls the spread of the distribution. A smaller sigma results in a narrower distribution, while a larger sigma leads to a wider distribution.

3. Perform the steps below with ”dat” dataframe which is just a sample data for you to observe how each plot function (3b through 3e) works. Notice that you need to have ggplot2 library installed on your system. Please refer slides how to install and import a library. Installation is done only once, but you need to import the library every time you need it by saying library(ggplot2). Then run the following commands for questions from 3a through 3e and observe how the plots are generated first. (20 points)
(a) dat <- data.frame(cond = factor(rep(c("A","B"), each=200)), rating = c(rnorm(200),rnorm(200, mean=.8)))
dat <- data.frame(cond = factor(rep(c("A","B"), each=200)), rating = c(rnorm(200),rnorm(200, mean=.8)))
(b) # Overlaid histograms

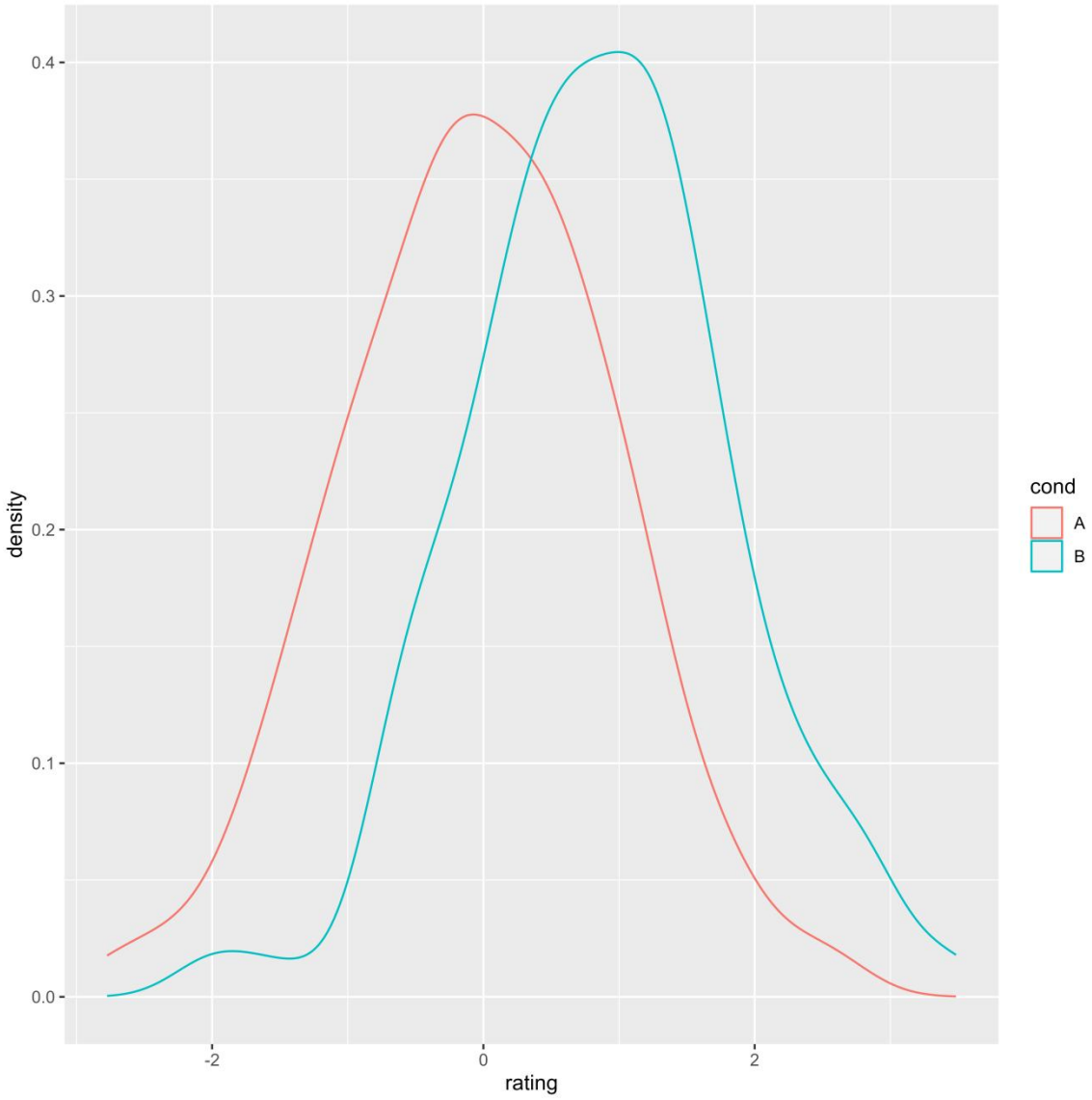
```
ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5, alpha=.5,
position="identity")
t1 = ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5, alpha=.5,
position="identity")
```
(c)
```
# Interleaved histograms
ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5,
position="dodge")
t2 = 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()
t3 = 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)
t4 = ggplot(dat, aes(x=rating, fill=cond)) + geom_density(alpha=.3)
```
(f) Read ”diabetes_train.csv” into a variable called diabetes and apply the same functions 3b through 3e for the mass attribute of diabetes and save the images. (Hint: instead of cond above, use the class attribute to color your groups. When you have fill option, your plots should show same type of chart for both groups in different colors on the same figure. Keep in mind that diabetes and dat are both DataFrames)
```
data.file <- file.path('data', 'diabetes_train.csv')
diabetes <- read.csv(data.file, header = TRUE, sep = ',')
p1 = ggplot(diabetes , aes(x=mass, fill=class)) + geom_histogram(binwidth=.5,
alpha=.5, position="identity")
p2 = ggplot(diabetes , aes(x=mass, fill=class)) + geom_histogram(binwidth=.5,
position="dodge")
p3 = ggplot(diabetes , aes(x=mass, colour=class)) + geom_density()
p4 = ggplot(diabetes , aes(x=mass, fill=class)) + geom_density(alpha=.3)
ggsave("histogram_3fb.png", plot = p1, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3fc.png", plot = p2, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3fd.png", plot = p3, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3fe.png", plot = p4, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3b.png", plot = t1, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3c.png", plot = t2, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3d.png", plot = t3, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3e.png", plot = t4, width = 8, height = 8, dpi = 1000)
```



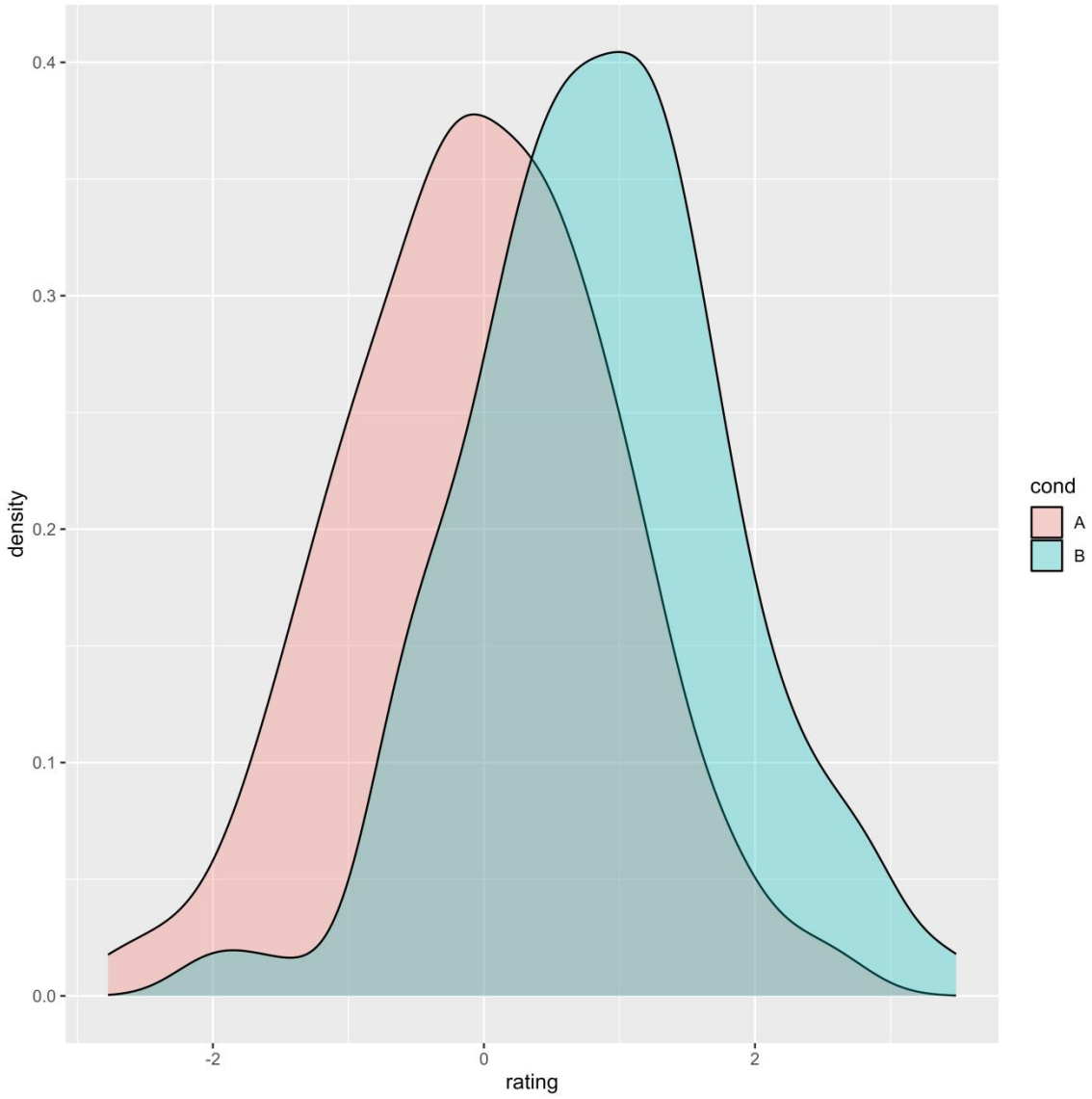
3b



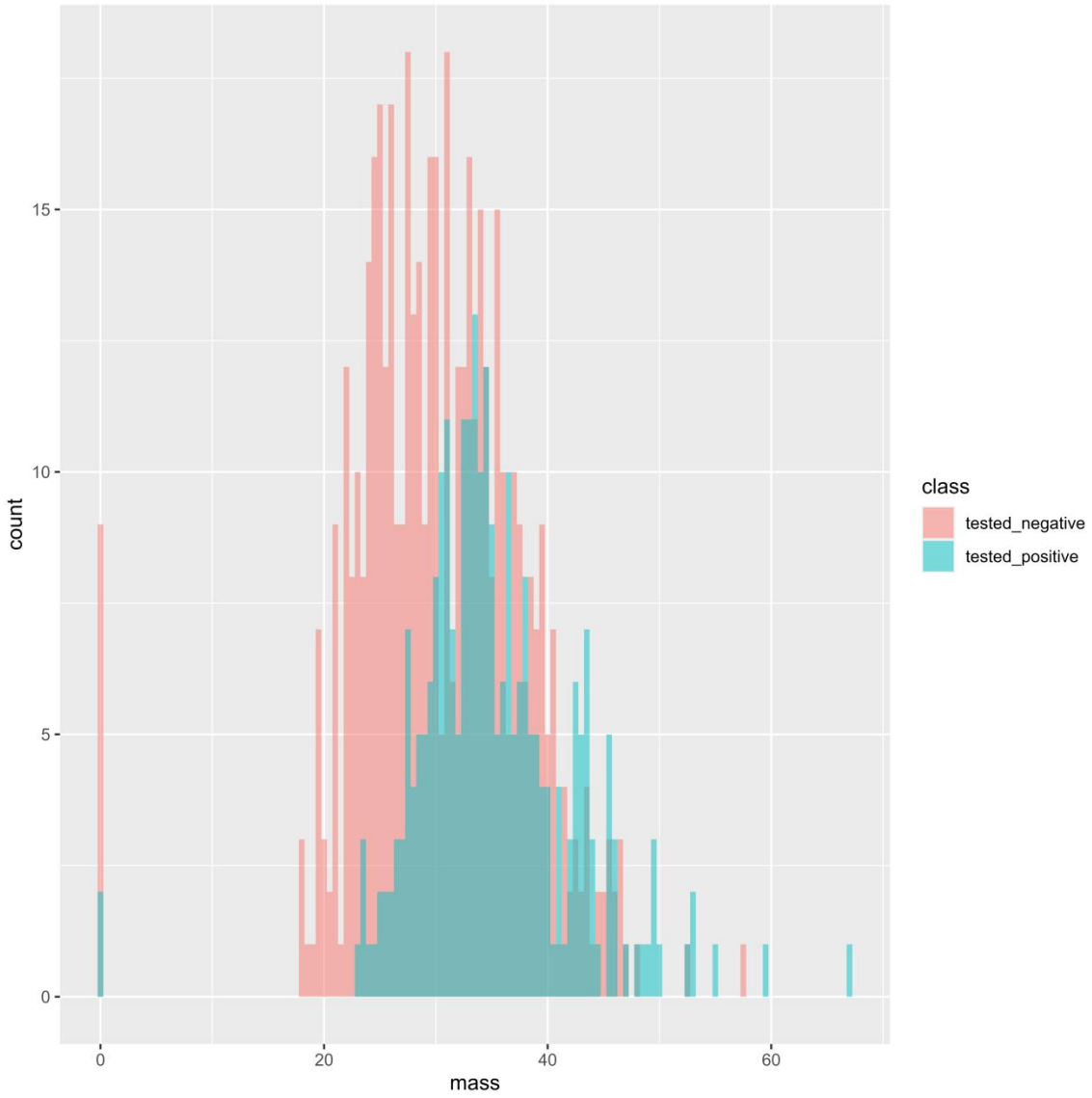
3c



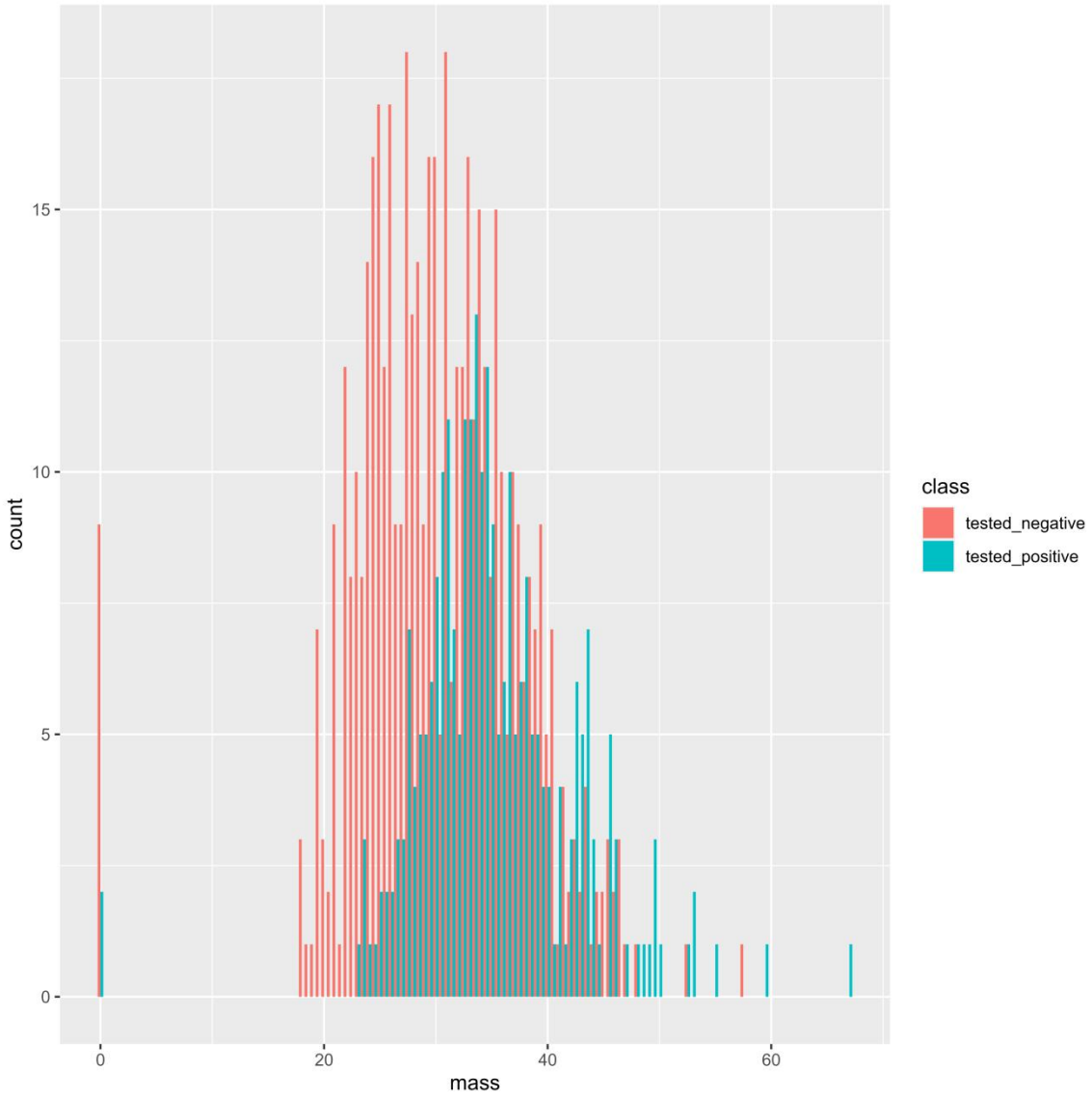
3d



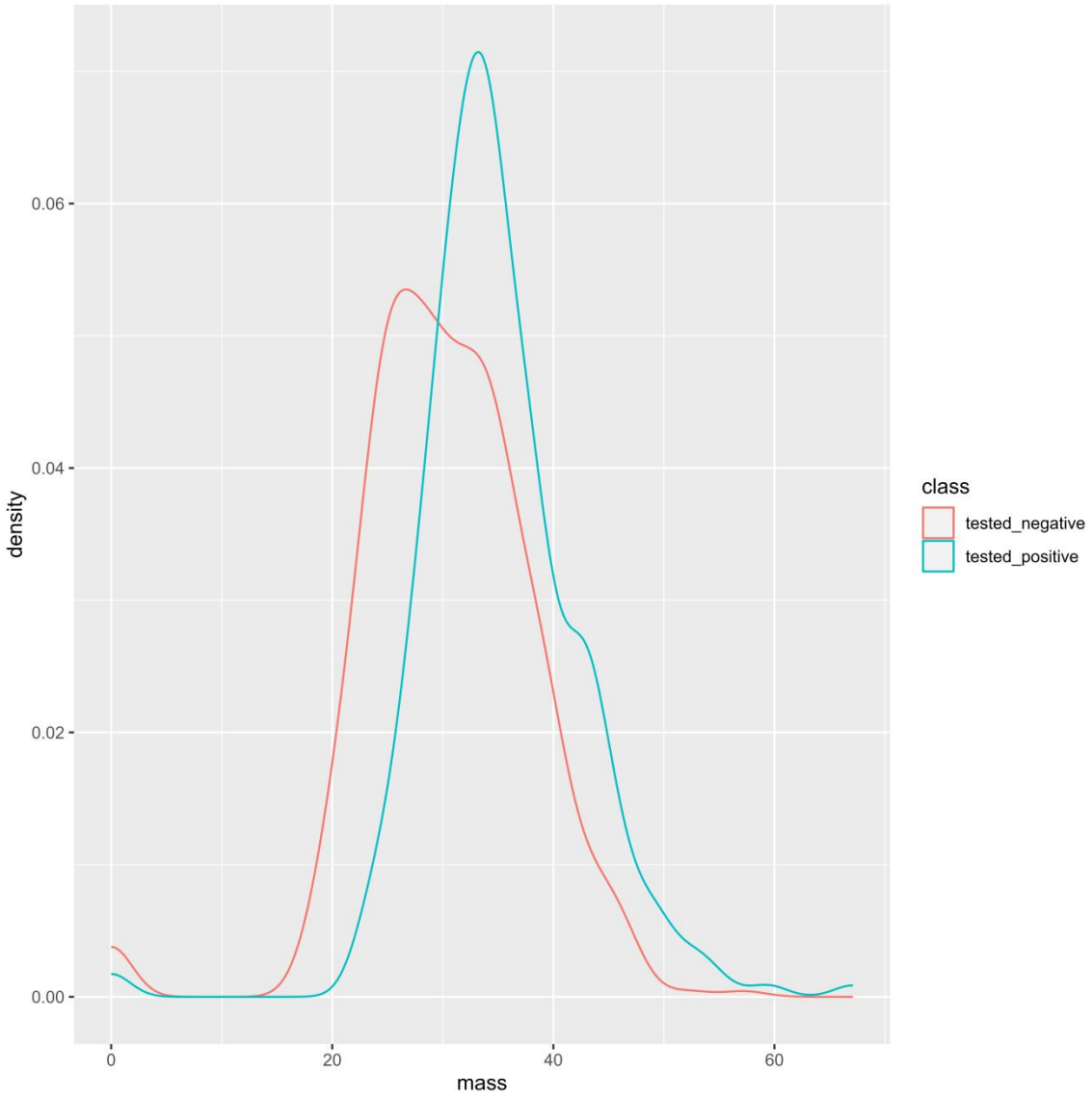
3e


count
15
10
5
0
0
20
40
60
mass
class
tested_negative
tested_positive

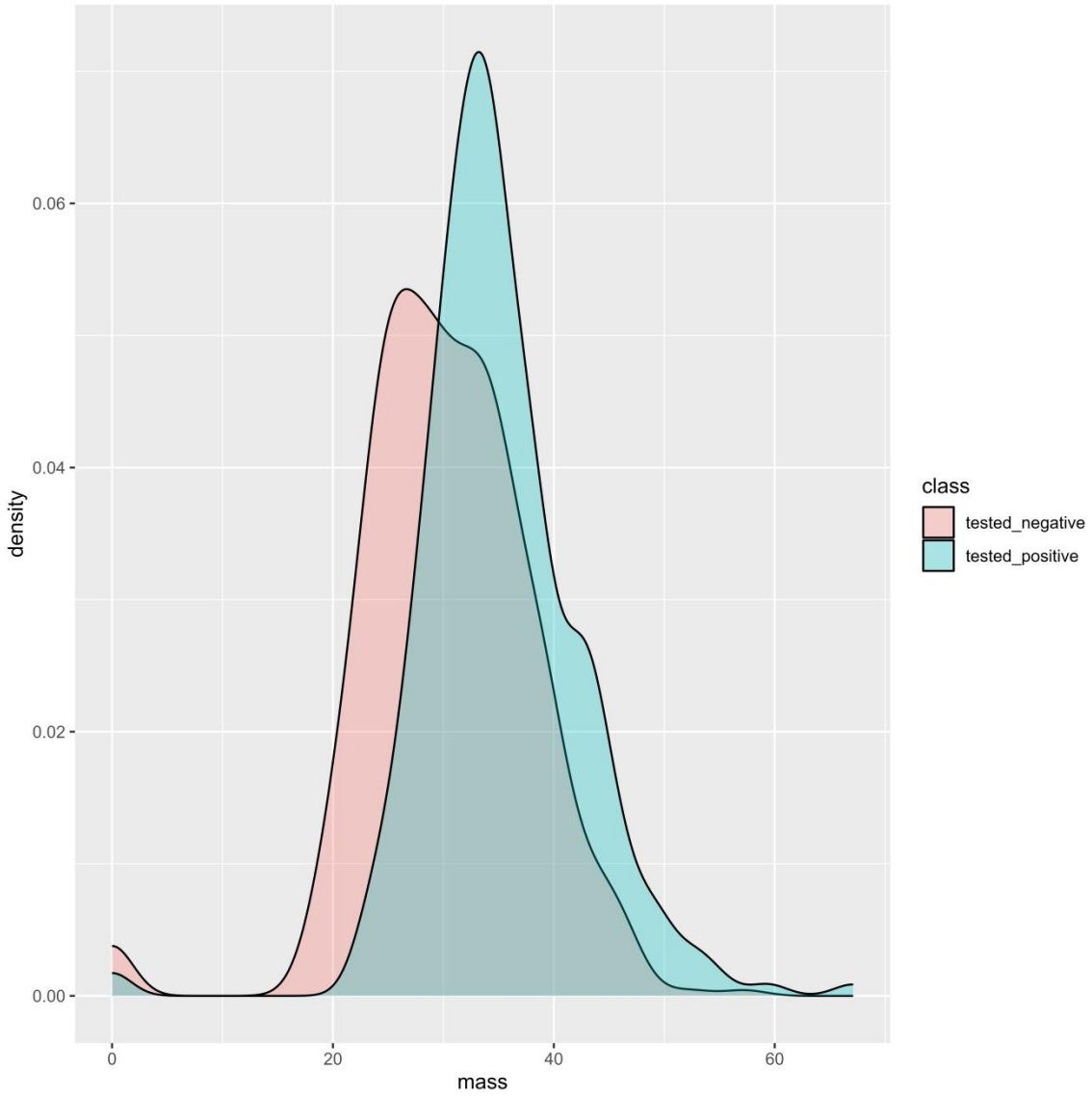
3fb


count
15
10
5
0
0
20
40
60
mass
class
tested_negative
tested_positive

3fc


0.06
0.04
0.02
0.00
density
0
20
40
60
mass
class
tested_negative
tested_positive

3fd



3fe

4. Read the titanic.csv file from DATA folder to a variable named passengers and perform the following steps and explain the operation very briefly (20 points):

data.file <- file.path('data', 'titanic.csv')
passengers <- read.csv(data.file, header = TRUE, sep = ',')
library(tidyr)
library(dplyr)
(a) passengers \%>\% drop_na() \%>\% summary()
First drops rows where any column contains a missing value from passengers. Then make the summary.
passengers %>% drop_na() %>% summary()

```
ggsave("histogram_3b.png", plot = t1, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3c.png", plot = t2, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3d.png", plot = t3, width = 8, height = 8, dpi = 1000)
ggsave("histogram_3e.png", plot = t4, width = 8, height = 8, dpi = 1000)

#4. Read the titanic.csv file from DATA folder to a variable named passengers and perform the following steps and
#explain the operation very briefly (20 points):
data.file <- file.path('data', 'titanic.csv')
passengers <- read.csv(data.file, header = TRUE, sep = ',')
#(a) passengers \%>\% drop_na() \%>\% summary()
#First drops rows where any column contains a missing value from passengers. Then make the summary.
```

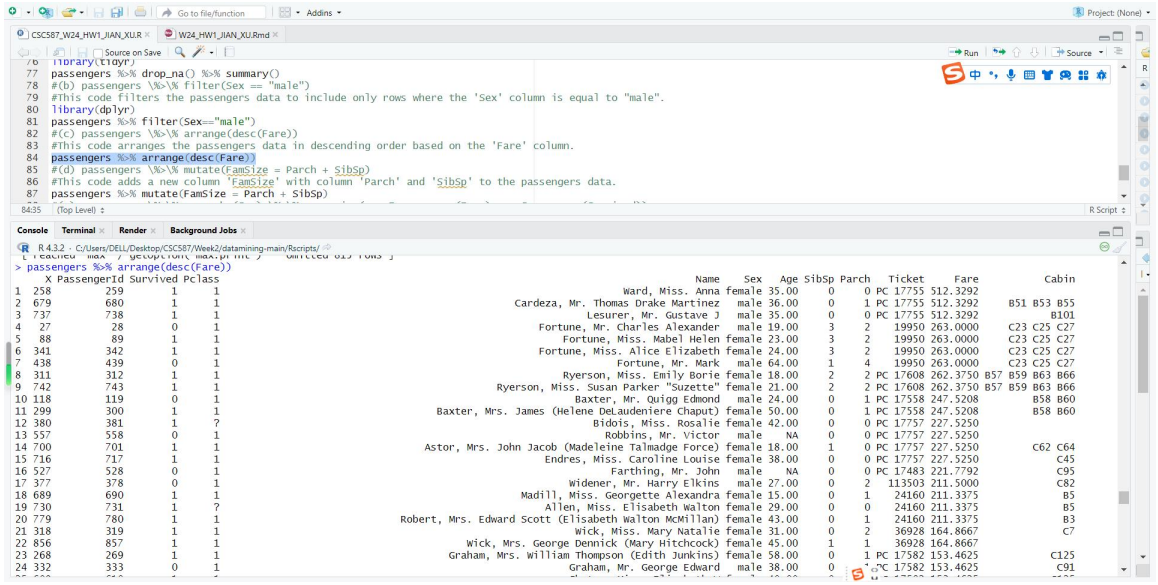
```
> ggsave("histogram_3fe.png", plot = p4, width = 8, height = 8, dpi = 1000)
> ggsave("histogram_3b.png", plot = t1, width = 8, height = 8, dpi = 1000)
> ggsave("histogram_3c.png", plot = t2, width = 8, height = 8, dpi = 1000)
> ggsave("histogram_3d.png", plot = t3, width = 8, height = 8, dpi = 1000)
> ggsave("histogram_3e.png", plot = t4, width = 8, height = 8, dpi = 1000)
> data.file <- file.path('data', 'titanic.csv')
> passengers <- read.csv(data.file, header = TRUE, sep = ',')
> #(a) passengers \%>\% drop_na() \%>\% summary()
> #First drops rows where any column contains a missing value from passengers. Then make the summary.
> library(tidyr)
> passengers %>% drop_na() %>% summary()
       X          PassengerId       Survived         Pclass             Name               Sex                 Age       
 Min.   :  0.0   Min.   :  1.0   Min.   :0.0000   Length:714         Length:714         Length:714         Min.   : 0.42  
 1st Qu.:221.2   1st Qu.:222.2   1st Qu.:0.0000   Class :character   Class :character   Class :character   1st Qu.:20.12  
 Median :444.0   Median :445.0   Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Median :28.00  
 Mean   :447.6   Mean   :448.6   Mean   :0.4062                                                           Mean   :29.70  
 3rd Qu.:676.8   3rd Qu.:677.8   3rd Qu.:1.0000                                                           3rd Qu.:38.00  
 Max.   :890.0   Max.   :891.0   Max.   :1.0000                                                           Max.   :80.00  
     SibSp           Parch           Ticket               Fare            Cabin              Embarked        
 Min.   :0.0000   Min.   :0.0000   Length:714         Min.   :  0.00   Length:714         Length:714        
 1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Qu.:  8.05   Class :character   Class :character  
 Median :0.0000   Median :0.0000   Mode  :character   Median : 15.74   Mode  :character   Mode  :character  
 Mean   :0.5126   Mean   :0.4314                      Mean   : 34.69                                        
 3rd Qu.:1.0000   3rd Qu.:1.0000                      3rd Qu.: 33.38                                        
 Max.   :5.0000   Max.   :6.0000                      Max.   :512.33                                        
> 
```

(b) passengers \%>\% filter(Sex == "male")
This code filters the passengers data to include only rows where the 'Sex' column is equal to "male".
passengers %>% filter(Sex=="male")

```
library(tidyr)
passengers %>% drop_na() %>% summary()
#(b) passengers \%>\% filter(Sex == "male")
#This code filters the passengers data to include only rows where the 'Sex' column is equal to "male".
library(dplyr)
passengers %>% filter(Sex=="male")
#(c) passengers \%>\% arrange(desc(Fare))
#This code arranges the passengers data in descending order based on the 'Fare' column.
passengers %>% arrange(desc(Fare))
#(d) passengers \%>\% mutate(FamSize = Parch + SibSp)
#This code adds a new column 'FamSize' with column 'Parch' and 'SibSp' to the passengers data.
passengers %>% mutate(FamSize = Parch + SibSp)
```

```
 [ reached 'max' / getOption("max.print") -- omitted 501 rows ]
> library(dplyr)
> passengers %>% filter(Sex=="male")
     X PassengerId Survived Pclass                                 Name  Sex   Age SibSp Parch           Ticket     Fare       Cabin Embarked
1    0           1        0      3              Braund, Mr. Owen Harris male 22.00     1     0        A/5 21171   7.2500                   S
2    4           5        0      3             Allen, Mr. William Henry male 35.00     0     0           373450   8.0500                   S
3    5           6        0      3                     Moran, Mr. James male    NA     0     0           330877   8.4583                   Q
4    6           7        0      1              McCarthy, Mr. Timothy J male 54.00     0     0            17463  51.8625         E46       S
5    7           8        0      3       Palsson, Master. Gosta Leonard male  2.00     3     1           349909  21.0750                   S
6   12          13        0      3     Saundercock, Mr. William Henry male 20.00     0     0        A/5. 2151   8.0500                   S
7   13          14        0      3        Andersson, Mr. Anders Johan male 39.00     1     5           347082  31.2750                   S
8   16          17        0      3               Rice, Master. Eugene male  2.00     4     1           382652  29.1250                   Q
9   17          18        1      2    Williams, Mr. Charles Eugene male    NA     0     0           244373  13.0000                   S
10  20          21        0      2              Fynney, Mr. Joseph J male 35.00     0     0           239865  26.0000                   S
11  21          22        1      2             Beesley, Mr. Lawrence male 34.00     0     0           248698  13.0000         D56       S
12  23          24        1      1       Sloper, Mr. William Thompson male 28.00     0     0           113788  35.5000          A6       S
13  26          27        0      3             Emir, Mr. Farred Chehab male    NA     0     0             2631   7.2250                   C
14  27          28        0      1      Fortune, Mr. Charles Alexander male 19.00     3     2            19950 263.0000 C23 C25 C27       S
15  29          30        0      3                  Todoroff, Mr. Lalio male    NA     0     0           349216   7.8958                   S
16  30          31        0      1           Uruchurtu, Don. Manuel E male 40.00     0     0         PC 17601  27.7208                   C
17  33          34        0      2              Wheadon, Mr. Edward H male 66.00     0     0        C.A. 24579  10.5000                   S
18  34          35        0      1             Meyer, Mr. Edgar Joseph male 28.00     1     0         PC 17604  82.1708                   C
19  35          36        0      1      Holverson, Mr. Alexander Oskar male 42.00     1     0           113789  52.0000                   S
20  36          37        1      3                     Mamee, Mr. Hanna male    NA     0     0             2677   7.2292                   C
21  37          38        0      ?           Cann, Mr. Ernest Charles male 21.00     0     0        A./5. 2152   8.0500                   S
22  42          43        0      3               Kraeff, Mr. Theodor male    NA     0     0           349253   7.8958                   S
```

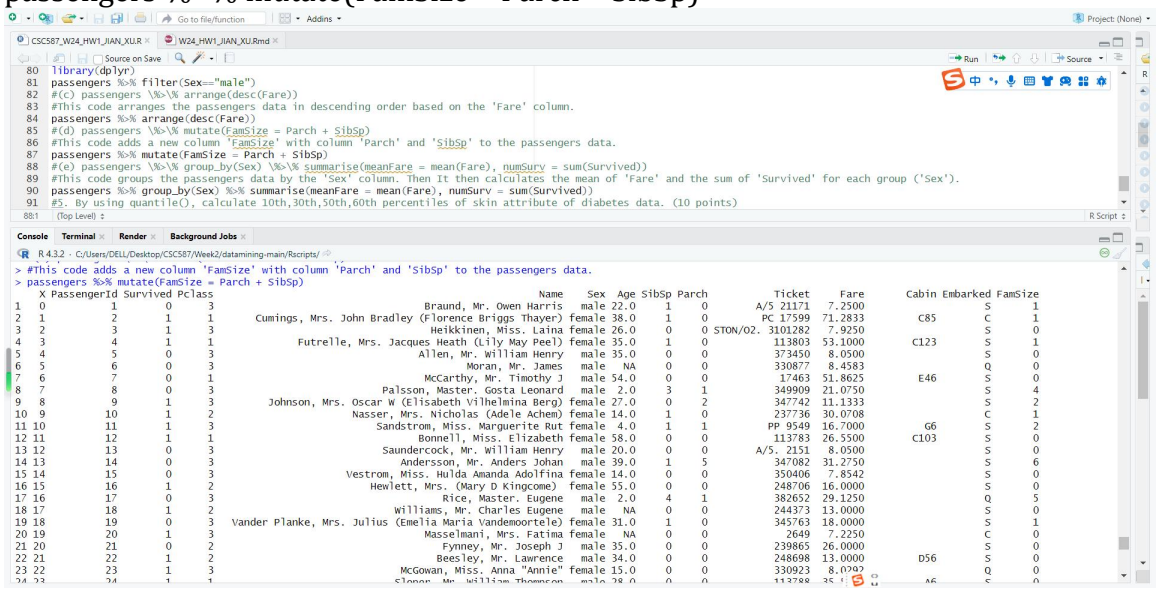
(c) passengers \%>\% arrange(desc(Fare))
This code arranges the passengers data in descending order based on the 'Fare' column.
passengers %>% arrange(desc(Fare))



```
passengers %>% drop_na() %>% summary()
#(b) passengers \%>\% filter(Sex == "male")
#This code filters the passengers data to include only rows where the 'Sex' column is equal to "male".
library(dplyr)
passengers %>% filter(Sex=="male")
#(c) passengers \%>\% arrange(desc(Fare))
#This code arranges the passengers data in descending order based on the 'Fare' column.
passengers %>% arrange(desc(Fare))
#(d) passengers \%>\% mutate(FamSize = Parch + SibSp)
#This code adds a new column 'FamSize' with column 'Parch' and 'SibSp' to the passengers data.
passengers %>% mutate(FamSize = Parch + SibSp)
```

```
> passengers %>% arrange(desc(Fare))
    X PassengerId Survived Pclass                                                 Name    Sex  Age SibSp Parch   Ticket     Fare           Cabin
1   258         259        1      1                                  Ward, Miss. Anna female 35.00     0     0 PC 17755 512.3292
2   679         680        1      1                Cardeza, Mr. Thomas Drake Martinez   male 36.00     0     1 PC 17755 512.3292     B51 B53 B55
3   737         738        1      1                            Lesurer, Mr. Gustave J   male 35.00     0     0 PC 17755 512.3292            B101
4    27          28        0      1                     Fortune, Mr. Charles Alexander   male 19.00     3     2    19950 263.0000     C23 C25 C27
5    88          89        1      1                        Fortune, Miss. Mabel Helen female 23.00     3     2    19950 263.0000     C23 C25 C27
6   341         342        1      1                      Fortune, Miss. Alice Elizabeth female 24.00     3     2    19950 263.0000     C23 C25 C27
7   438         439        0      1                                Fortune, Mr. Mark   male 64.00     1     4    19950 263.0000     C23 C25 C27
8   311         312        1      1                         Ryerson, Miss. Emily Borie female 18.00     2     2 PC 17608 262.3750 B57 B59 B63 B66
9   742         743        1      1               Ryerson, Miss. Susan Parker "Suzette" female 21.00     2     2 PC 17608 262.3750 B57 B59 B63 B66
10  118         119        0      1                          Baxter, Mr. Quigg Edmond   male 24.00     0     1 PC 17558 247.5208            B58 B60
11  299         300        1      1          Baxter, Mrs. James (Helene DeLaudeniere Chaput) female 50.00     0     1 PC 17558 247.5208            B58 B60
12  380         381        1      ?                                Bidois, Miss. Rosalie female 42.00     0     0 PC 17757 227.5250
13  557         558        0      1                                  Robbins, Mr. Victor   male    NA     0     0 PC 17757 227.5250
14  700         701        1      1          Astor, Mrs. John Jacob (Madeleine Talmadge Force) female 18.00     1     0 PC 17757 227.5250          C62 C64
15  716         717        1      1                       Endres, Miss. Caroline Louise female 38.00     0     0 PC 17757 227.5250              C45
16  527         528        0      1                                 Farthing, Mr. John   male    NA     0     0 PC 17483 221.7792              C95
17  377         378        0      1                             Widener, Mr. Harry Elkins   male 27.00     0     2   113503 211.5000              C82
18  689         690        1      1                          Madill, Miss. Georgette Alexandra female 15.00     0     1    24160 211.3375               B5
19  730         731        1      ?                                Allen, Miss. Elisabeth Walton female 29.00     0     0    24160 211.3375               B5
20  779         780        1      1          Robert, Mrs. Edward Scott (Elisabeth Walton McMillan) female 43.00     0     1    24160 211.3375               B3
21  318         319        1      1                                  Wick, Miss. Mary Natalie female 31.00     0     2    36928 164.8667               C7
22  856         857        1      1                          Wick, Mrs. George Dennick (Mary Hitchcock) female 45.00     1     1    36928 164.8667
23  268         269        1      1                       Graham, Mrs. William Thompson (Edith Junkins) female 58.00     0     1 PC 17582 153.4625             C125
24  332         333        0      1                                 Graham, Mr. George Edward   male 38.00     0     1 PC 17582 153.4625              C91
```

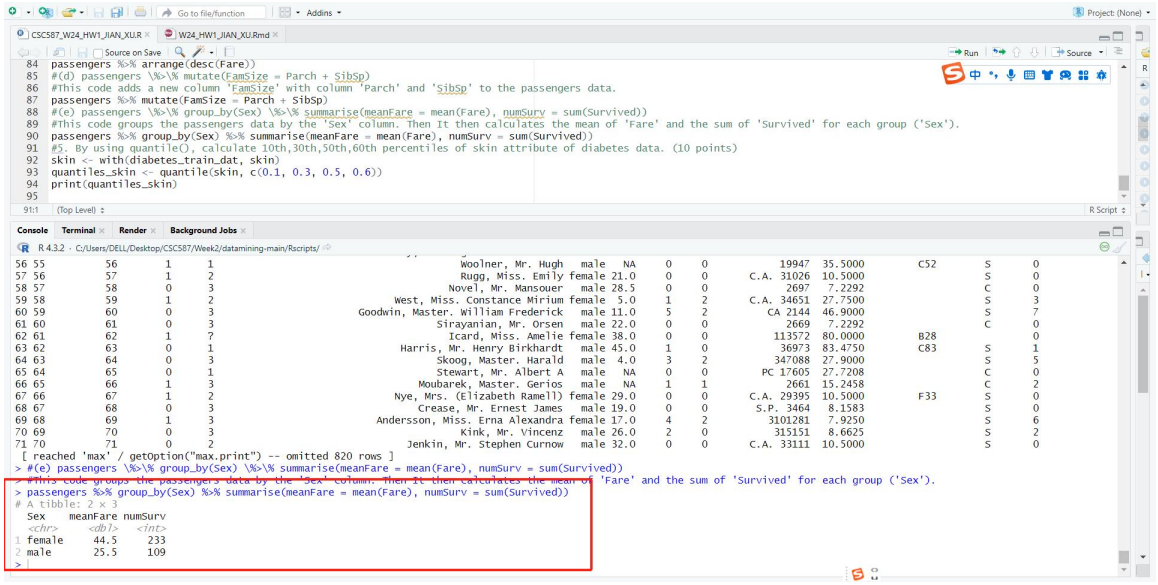
(d) passengers \%>\% mutate(FamSize = Parch + SibSp)
This code adds a new column 'FamSize' with the value of column 'Parch' plus the value of column 'SibSp' to the passengers data.
passengers %>% mutate(FamSize = Parch + SibSp)



```
library(dplyr)
passengers %>% filter(Sex=="male")
#(c) passengers \%>\% arrange(desc(Fare))
#This code arranges the passengers data in descending order based on the 'Fare' column.
passengers %>% arrange(desc(Fare))
#(d) passengers \%>\% mutate(FamSize = Parch + SibSp)
#This code adds a new column 'FamSize' with column 'Parch' and 'SibSp' to the passengers data.
passengers %>% mutate(FamSize = Parch + SibSp)
#(e) passengers \%>\% group_by(Sex) \%>\% summarise(meanFare = mean(Fare), numSurv = sum(Survived))
#This code groups the passengers data by the 'Sex' column. Then It then calculates the mean of 'Fare' and the sum of 'Survived' for each group ('Sex').
passengers %>% group_by(Sex) %>% summarise(meanFare = mean(Fare), numSurv = sum(Survived))
#5. By using quantile(), calculate 10th,30th,50th,60th percentiles of skin attribute of diabetes data. (10 points)
```

```
> #This code adds a new column 'FamSize' with column 'Parch' and 'SibSp' to the passengers data.
> passengers %>% mutate(FamSize = Parch + SibSp)
    X PassengerId Survived Pclass                                                 Name    Sex  Age SibSp Parch           Ticket     Fare        Cabin Embarked FamSize
1   0           1        0      3                                Braund, Mr. Owen Harris   male 22.0     1     0        A/5 21171   7.2500                S        1
2   1           2        1      1      Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.0     1     0         PC 17599  71.2833          C85        C        1
3   2           3        1      3                                 Heikkinen, Miss. Laina female 26.0     0     0 STON/O2. 3101282   7.9250                S        0
4   3           4        1      1           Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0     1     0           113803  53.1000         C123        S        1
5   4           5        0      3                               Allen, Mr. William Henry   male 35.0     0     0           373450   8.0500                S        0
6   5           6        0      3                                       Moran, Mr. James   male   NA     0     0           330877   8.4583                Q        0
7   6           7        0      1                                McCarthy, Mr. Timothy J   male 54.0     0     0            17463  51.8625          E46        S        0
8   7           8        0      3                         Palsson, Master. Gosta Leonard   male  2.0     3     1           349909  21.0750                S        4
9   8           9        1      3      Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female 27.0     0     2           347742  11.1333                S        2
10  9          10        1      2                    Nasser, Mrs. Nicholas (Adele Achem) female 14.0     1     0           237736  30.0708                C        1
11 10          11        1      3                          Sandstrom, Miss. Marguerite Rut female  4.0     1     1          PP 9549  16.7000          G6        S        2
12 11          12        1      1                                 Bonnell, Miss. Elizabeth female 58.0     0     0           113783  26.5500         C103        S        0
13 12          13        0      3                           Saundercock, Mr. William Henry   male 20.0     0     0        A/5. 2151   8.0500                S        0
14 13          14        0      3                              Andersson, Mr. Anders Johan   male 39.0     1     5           347082  31.2750                S        6
15 14          15        0      3                  Vestrom, Miss. Hulda Amanda Adolfina female 14.0     0     0           350406   7.8542                S        0
16 15          16        1      2                      Hewlett, Mrs. (Mary D Kingcome)  female 55.0     0     0           248706  16.0000                S        0
17 16          17        0      3                                  Rice, Master. Eugene   male  2.0     4     1           382652  29.1250                Q        5
18 17          18        1      2                          Williams, Mr. Charles Eugene   male   NA     0     0           244373  13.0000                S        0
19 18          19        0      3 Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) female 31.0     1     0           345763  18.0000                S        1
20 19          20        1      3                           Masselmani, Mrs. Fatima female   NA     0     0             2649   7.2250                C        0
21 20          21        0      2                               Fynney, Mr. Joseph J   male 35.0     0     0           239865  26.0000                S        0
22 21          22        1      2                               Beesley, Mr. Lawrence   male 34.0     0     0           248698  13.0000          D56        S        0
23 22          23        1      3                         McGowan, Miss. Anna "Annie" female 15.0     0     0           330923   8.0292                Q        0
```

(e) passengers \%>\% group_by(Sex) \%>\% summarise(meanFare = mean(Fare), numSurv = sum(Survived))
This code groups the passengers data by the 'Sex' column. Then It then calculates the mean of 'Fare' and the sum of 'Survived' for each group ('Sex').
passengers %>% group_by(Sex) %>% summarise(meanFare = mean(Fare), numSurv = sum(Survived))



5. By using quantile(), calculate 10th,30th,50th,60th percentiles of skin attribute of diabetes data. (10 points)
skin <- with(diabetes_train_dat, skin)
quantiles_skin <- quantile(skin, c(0.1, 0.3, 0.5, 0.6))
print(quantiles_skin)