CSC 587-W1

Homework 1

Mike Turley

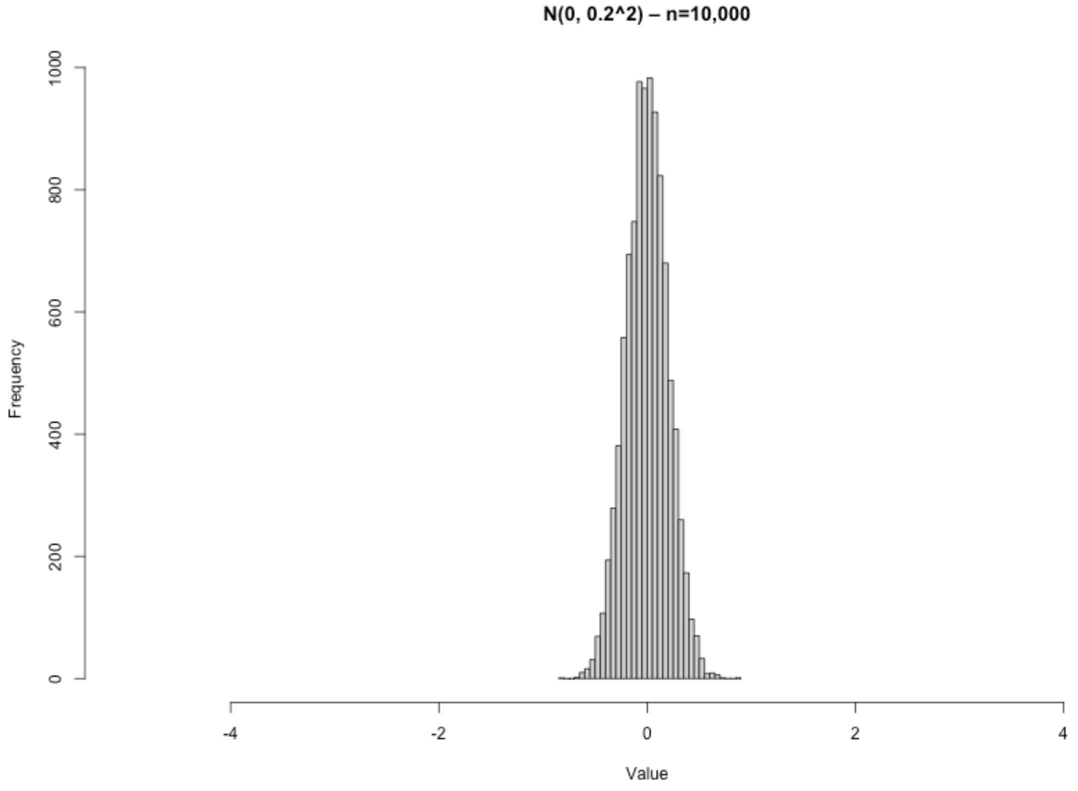
1.

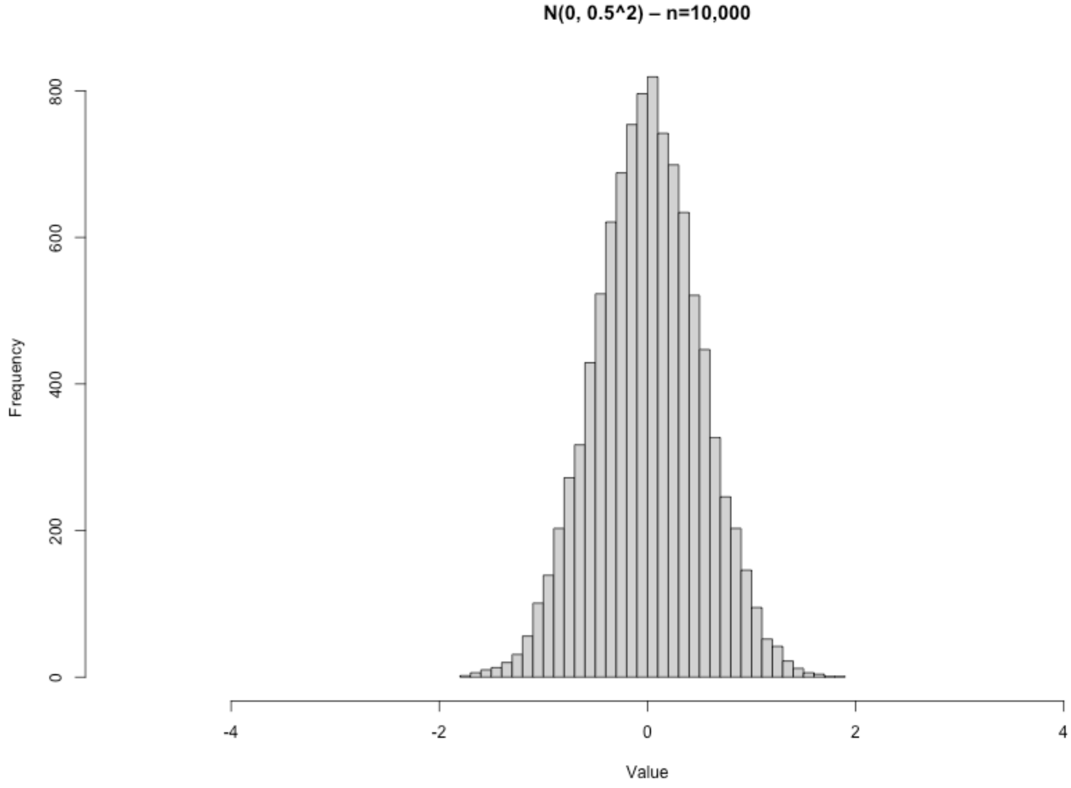
1**a**) Loaded Su_raw_matrix.txt into R using read.delim, confirming that the header included Liver_2.CEL.

1**b**) Calculated the mean and standard deviation of Liver_2.CEL to summarize its central tendency and variability.

1**c**) Computed column-wise means and sums for all numeric variables, giving an overview of overall expression levels across samples.

2)





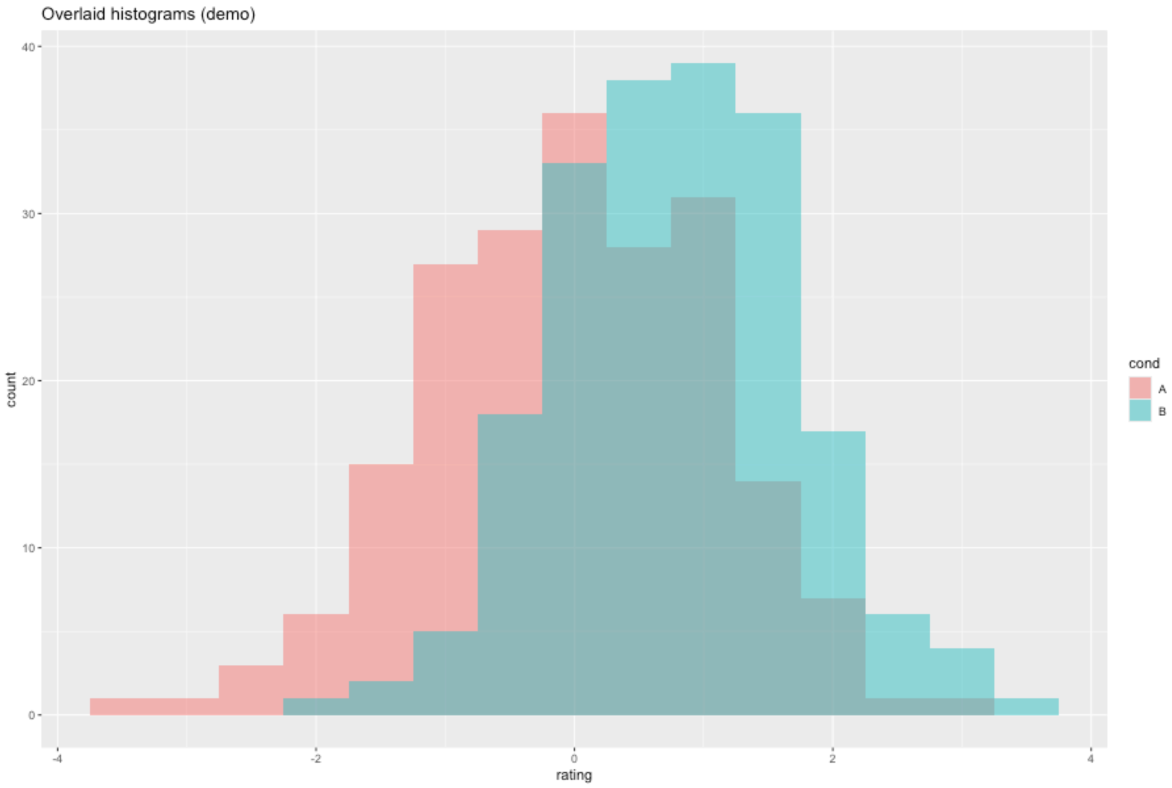
Histogram differences between rnorm with σ = 0.2 vs 0.5

Both sets are centered at 0 (same mean), but **σ = 0.5** spreads mass more widely than **σ = 0.2**. With identical x-limits (i.e., −5 to 5), the **σ = 0.2** histogram shows a taller, narrower shape; **σ = 0.5** shows a lower, wider shape. This shows a basic property of the Normal distribution: increasing the standard deviation increases dispersion while keeping the peak at the mean.

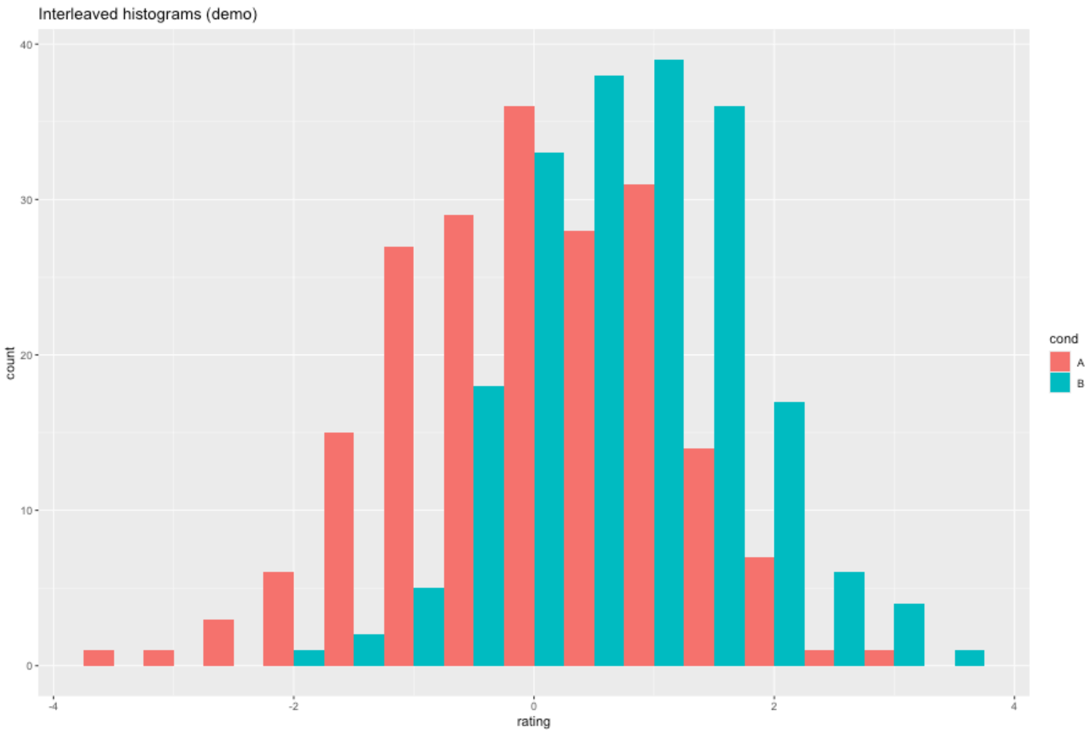
3.

3a) Create the sample dataframe

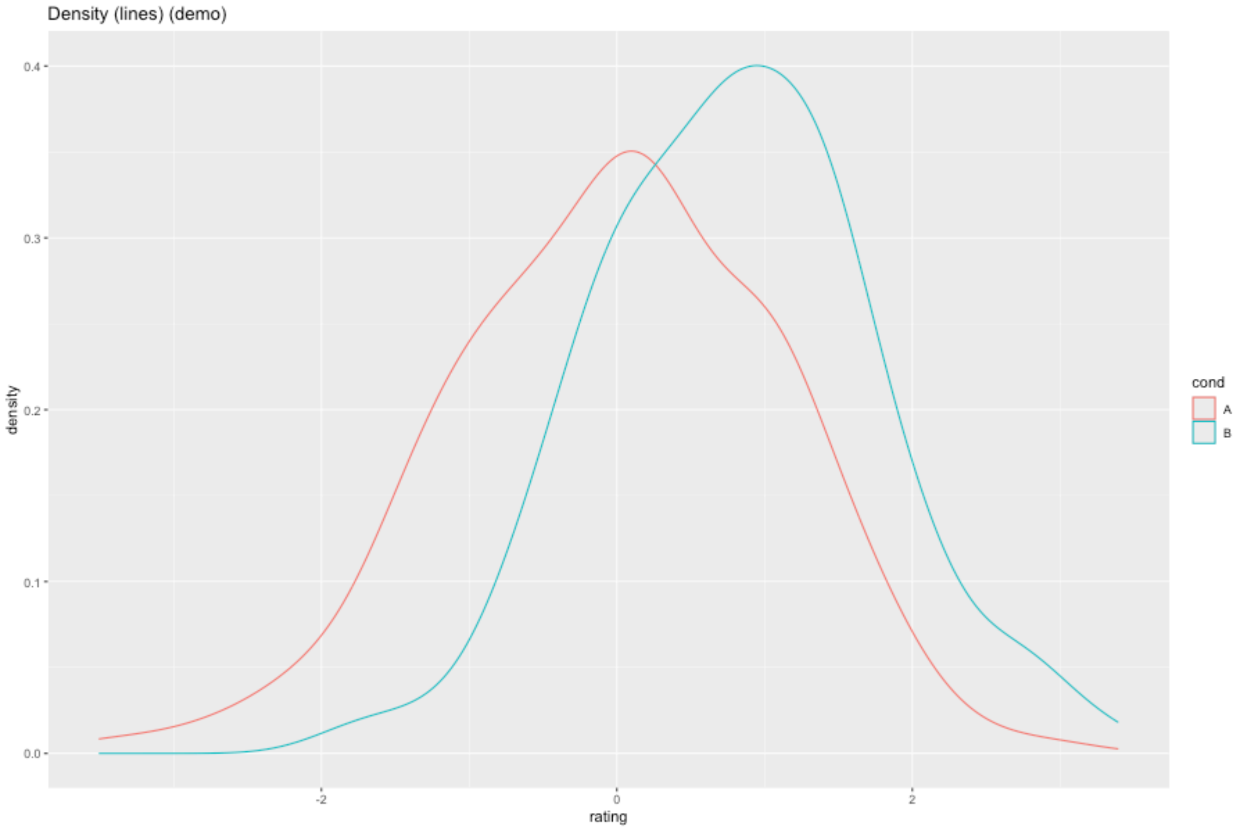
3b)


Overlaid histograms (demo)
count
40
30
20
10
0
-4
-2
0
2
4
rating
cond
A
B

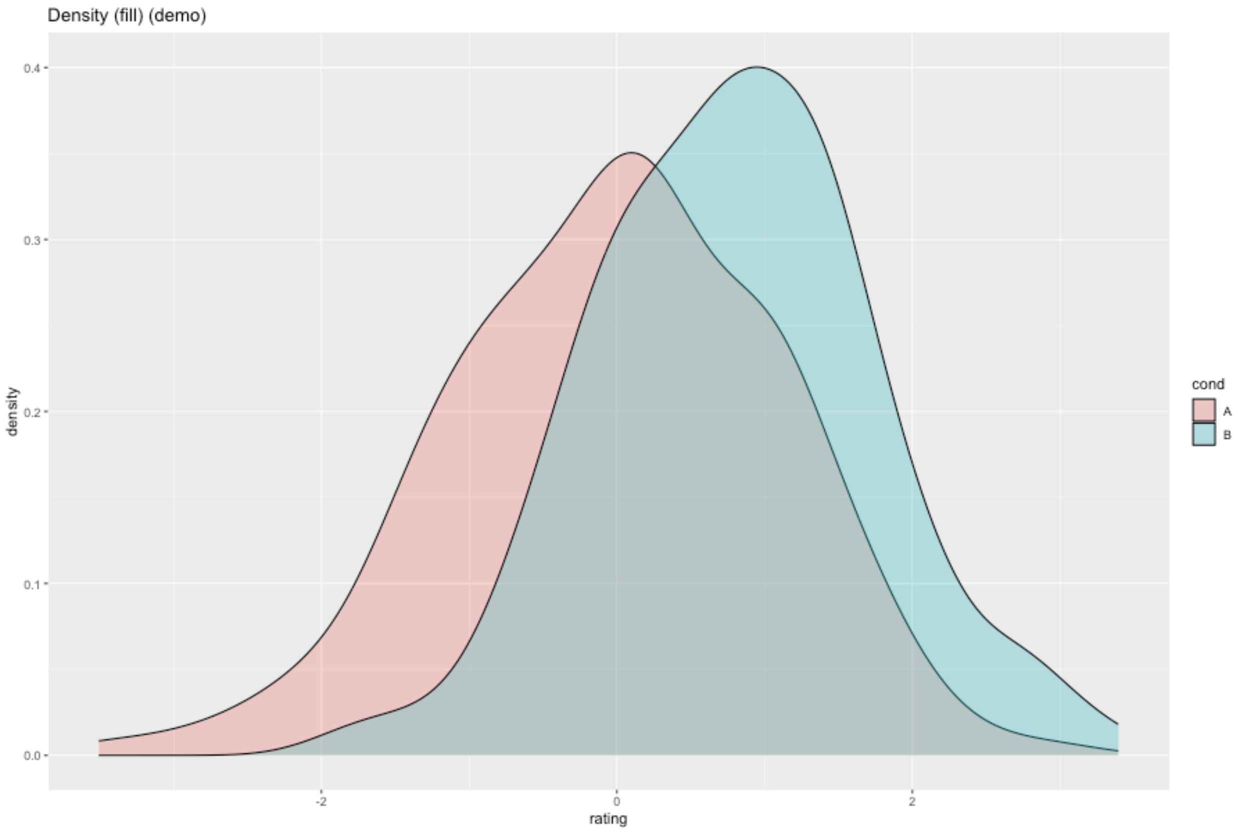
3c)


Interleaved histograms (demo)
count
40
30
20
10
0
-4
-2
0
2
4
rating
cond
A
B

3d)



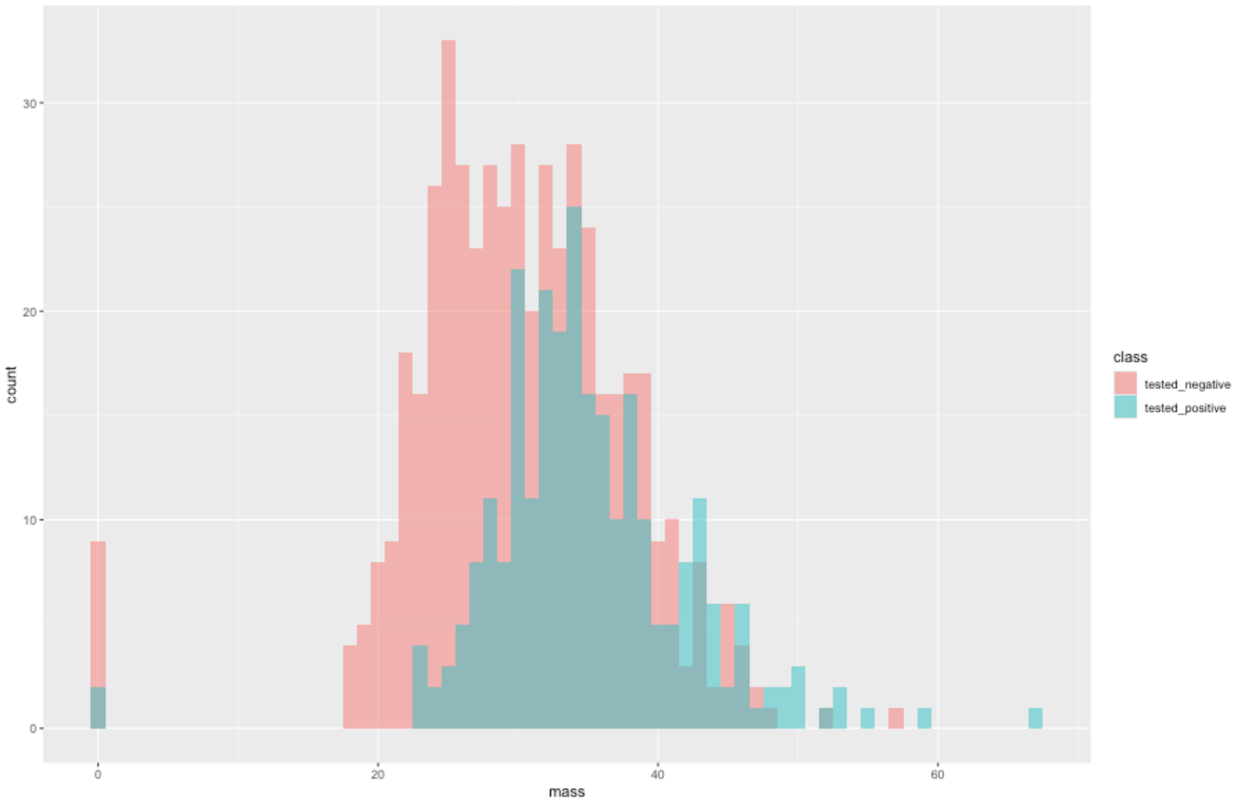
3e)



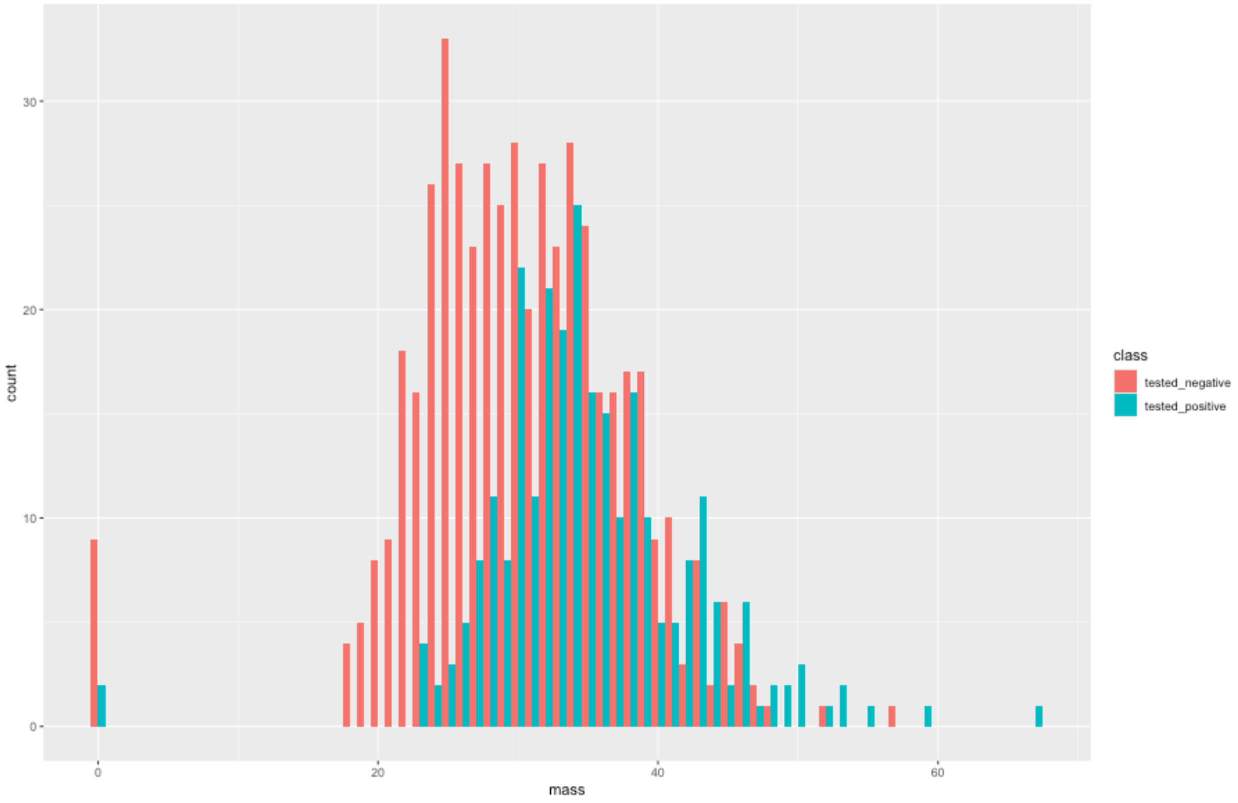
3f) Diabetes mass by class

Interleaved separates bars by class; overlaid stacks them in the same x-bins (transparency helps).

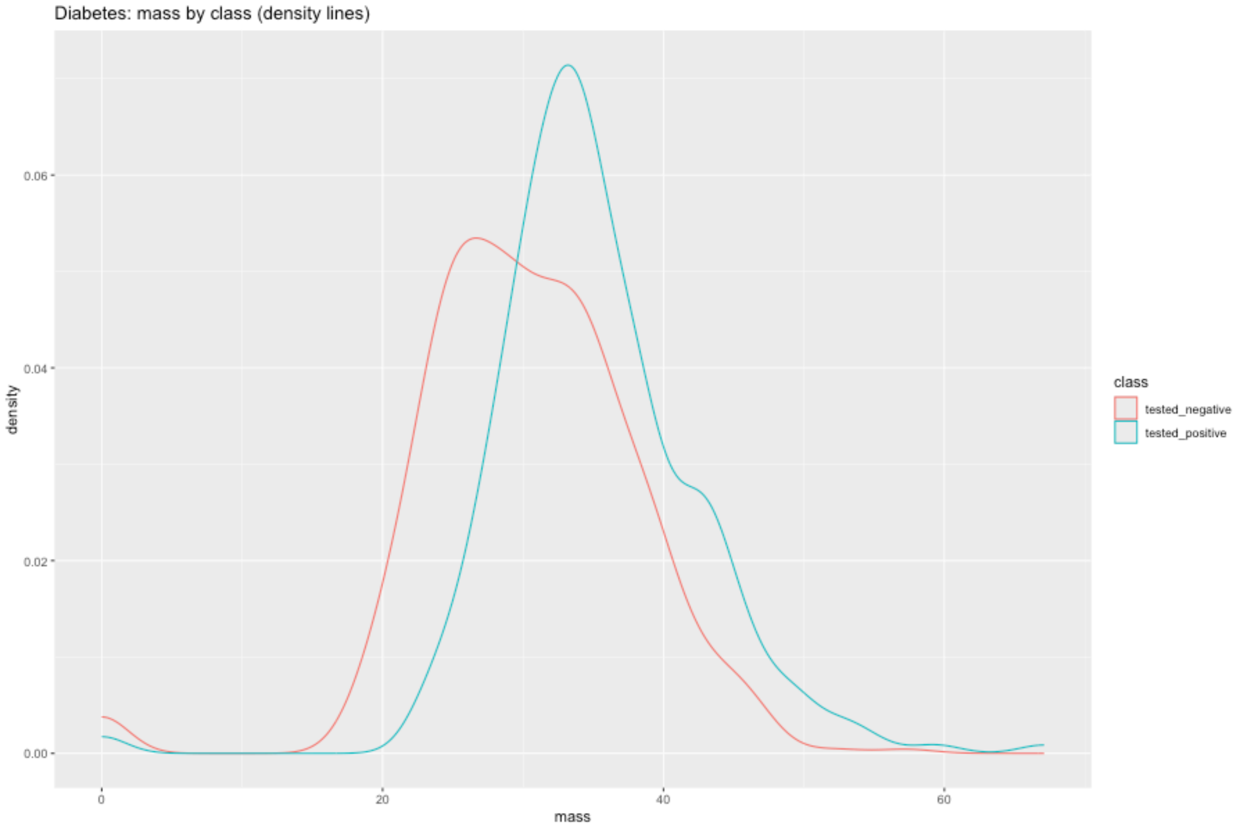
Diabetes: mass by class (overlaid)

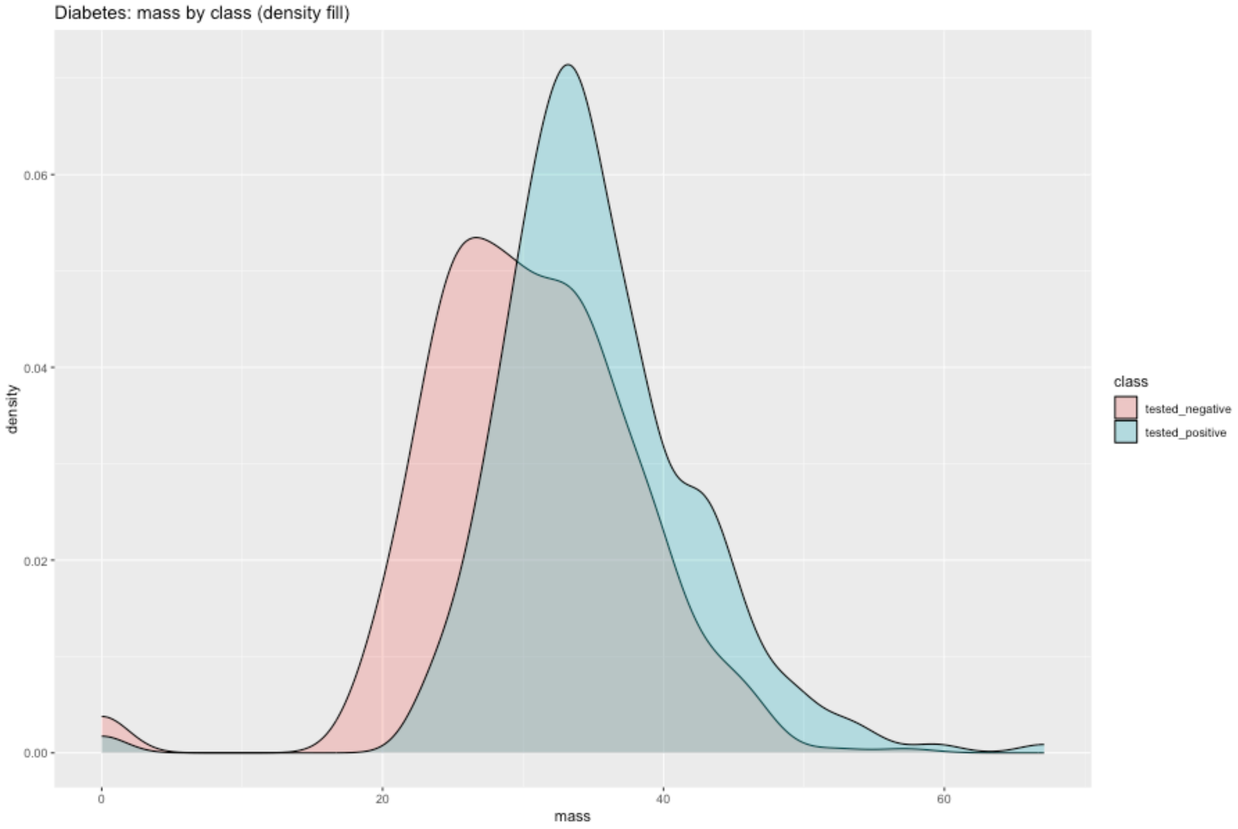


Diabetes: mass by class (interleaved)



For the density the lines emphasize distribution shape; filled density with alpha visually conveys overlap and separation between classes.


Diabetes: mass by class (density lines)


Diabetes: mass by class (density fill)

4.

4a)`drop_na() %>% summary()`: removes rows with missing values, then prints univariate summaries, ensuring stats aren't biased by NAs.

4b)`filter(Sex == "male")`: subsets to male passengers.

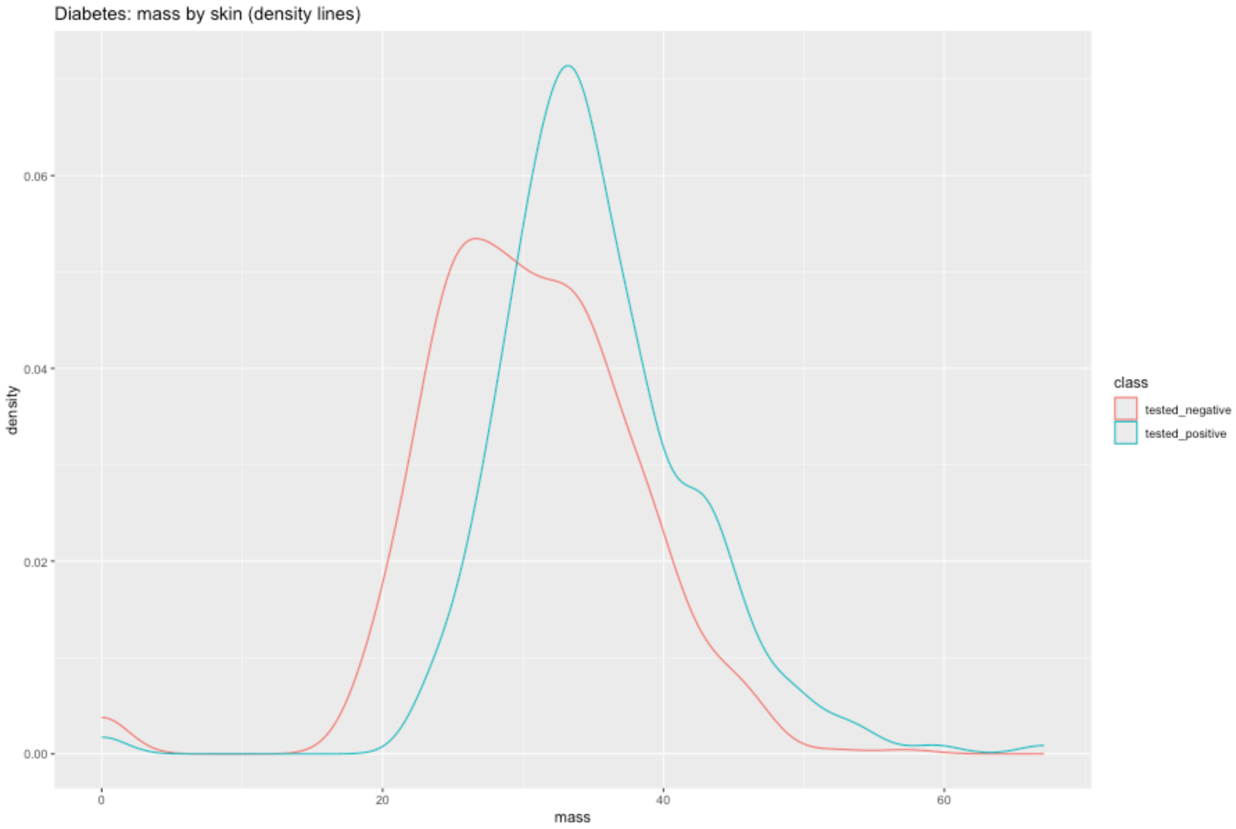
4c) `arrange(desc(Fare))`: sorts passengers by descending fare (highest first).

4d) `mutate(FamSize = Parch + SibSp)`: creates Family Size as sum of parents/ children and siblings/spouses.

4e) `group_by(Sex) %>% summarise(...)`: aggregates by sex to report mean fare and number of survivors.

5.

Used the `quantile()` function in R to compute the **10th, 30th, 50th, and 60th percentiles** of the `skin` variable in the diabetes dataset. These values highlight key points in the distribution and help us understand how skin thickness is spread across patients.


Diabetes: mass by skin (density lines)
density
mass
class
tested_negative
tested_positive