STAT 210
Applied Statistics and Data Analysis:
Homework 3

Due on Oct. 5/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

Question 1

For this question use the data set `penguins`, which is available in the `palmerpenguins` library. This data set was introduced in the second problem list and has four physical measurements for three species of penguins studied in Antartica. The species are `Adelie`, `Chinstrap` and `Gentoo`, and there are a total of 344 subjects, some with missing values. The penguins were observed in three different islands, `Biscoe`, `Dream` and `Togersen`. You can get more information looking at the help for this data set.

a) Find out

- how many subjects belong to each species,
- how many were observed in each island,
- how many missing values are there.

```
library(palmerpenguins)
```

```
##
## Attaching package: 'palmerpenguins'

## The following objects are masked from 'package:datasets':
##
##     penguins, penguins_raw
```

```
table(penguins$species) #extract species to a table
```

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

```
table(penguins$island) # extract islands to a table
```

```
##
##    Biscoe     Dream Torgersen
##       168       124        52
```

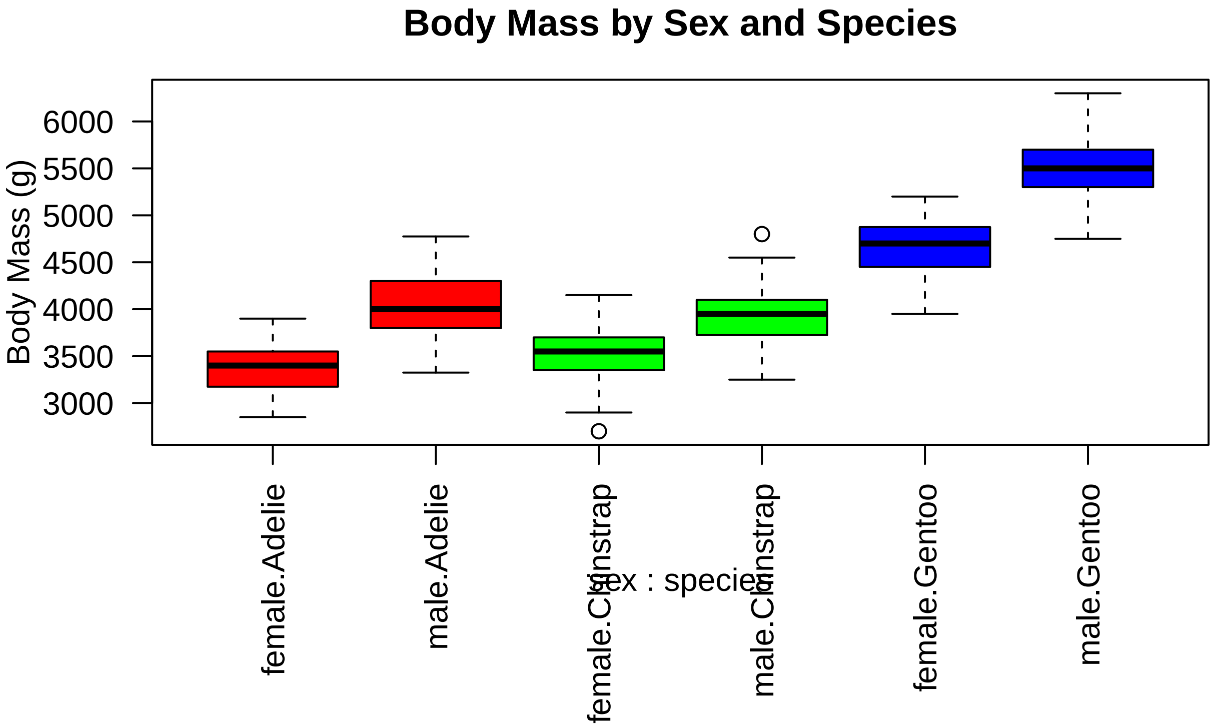
```
sum(is.na(penguins)) # find the sum of empty values in penguins
```

```
## [1] 19
```

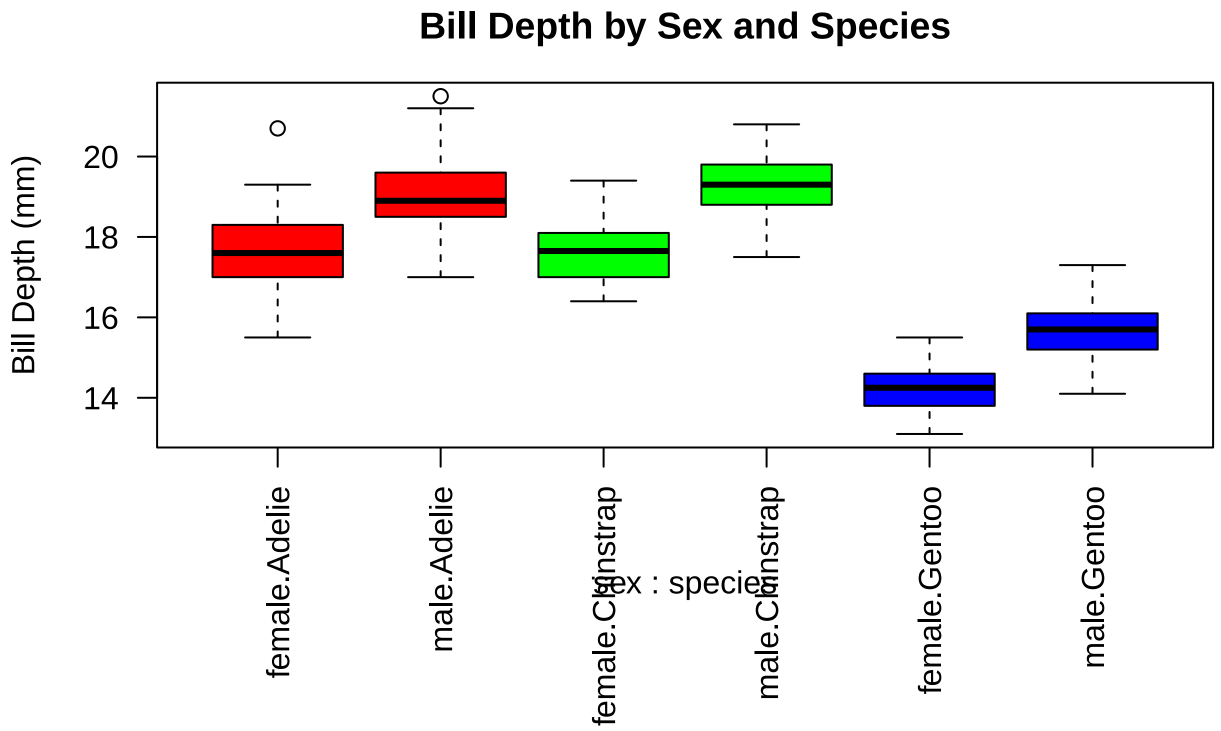
b) On a single plotting window, create boxplots of body mass grouped by both sex **and** species. This should result in six boxplots, one for each combination of sex and species. You can use the `boxplot` function, which allows you to specify a formula to define the variables for the plot. In the formula, use either `sex:species` or `sex + species` on the right-hand side to indicate grouping. Color the boxes according to species to distinguish them visually. Then, comment on what you observe in the resulting plot. Repeat the same process for bill depth, and again provide comments based on your observations.

```
#species_colors <- c("Adelie" = "red",
#                     "Chinstrap" = "blue",
  #                   "Gentoo" = "green")
penguins_f = na.omit(penguins)

par(mar = c(10, 4, 3, 1))
boxplot(body_mass_g ~ sex + species,
        data = penguins_f,
        col = rep(c("red", "green", "blue"), each = 2),
        main = "Body Mass by Sex and Species",
        ylab = "Body Mass (g)",
        las = 2)
```



```r
boxplot(bill_depth_mm ~ sex + species,
        data = penguins_f,
        col = rep(c("red", "green", "blue"), each = 2),
        main = "Bill Depth by Sex and Species",
        ylab = "Bill Depth (mm)",
        las = 2)
```



From the first plot of body mass by sex and species. we can see that generally males are heavier than females in all species. with the Gentoo having the highest weight, and the Adelie being the smallest.

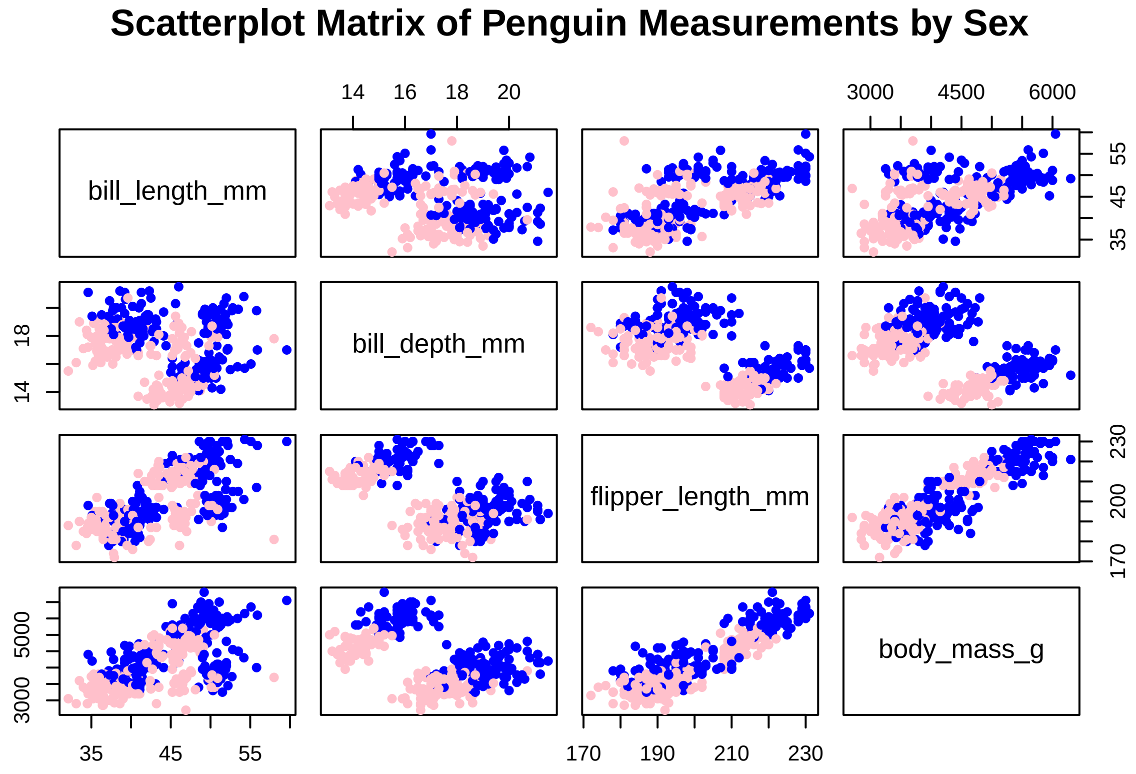
and using the 2nd boxplot we see that the bill depth of males is also larger than that of females, with the gentoo's having the lowest bill depth and the Adelie having the largest depth.

We can also see that the Adelie and Chinstrap have pretty similar bodymass and bill depth. with the Chinstrap having more variablility in mass and the Adelie having more variablity in bill depth.

c) Create a scatterplot matrix with plots of the four numerical variables in penguins. Color the plots according to `sex`. Comment on what you observe.

```r
#take out the columns of variables
numeric_vars <- penguins_f[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm","body_mass_g")]
#store the colors based on sex
sex_colors <- ifelse(penguins_f$sex == "male", "blue", "pink")

#plot the scatterplot matrix, pch is the plotting character (16 is a dot)
pairs(numeric_vars,
      col = sex_colors,
      pch = 16,
      main = "Scatterplot Matrix of Penguin Measurements by Sex")
```



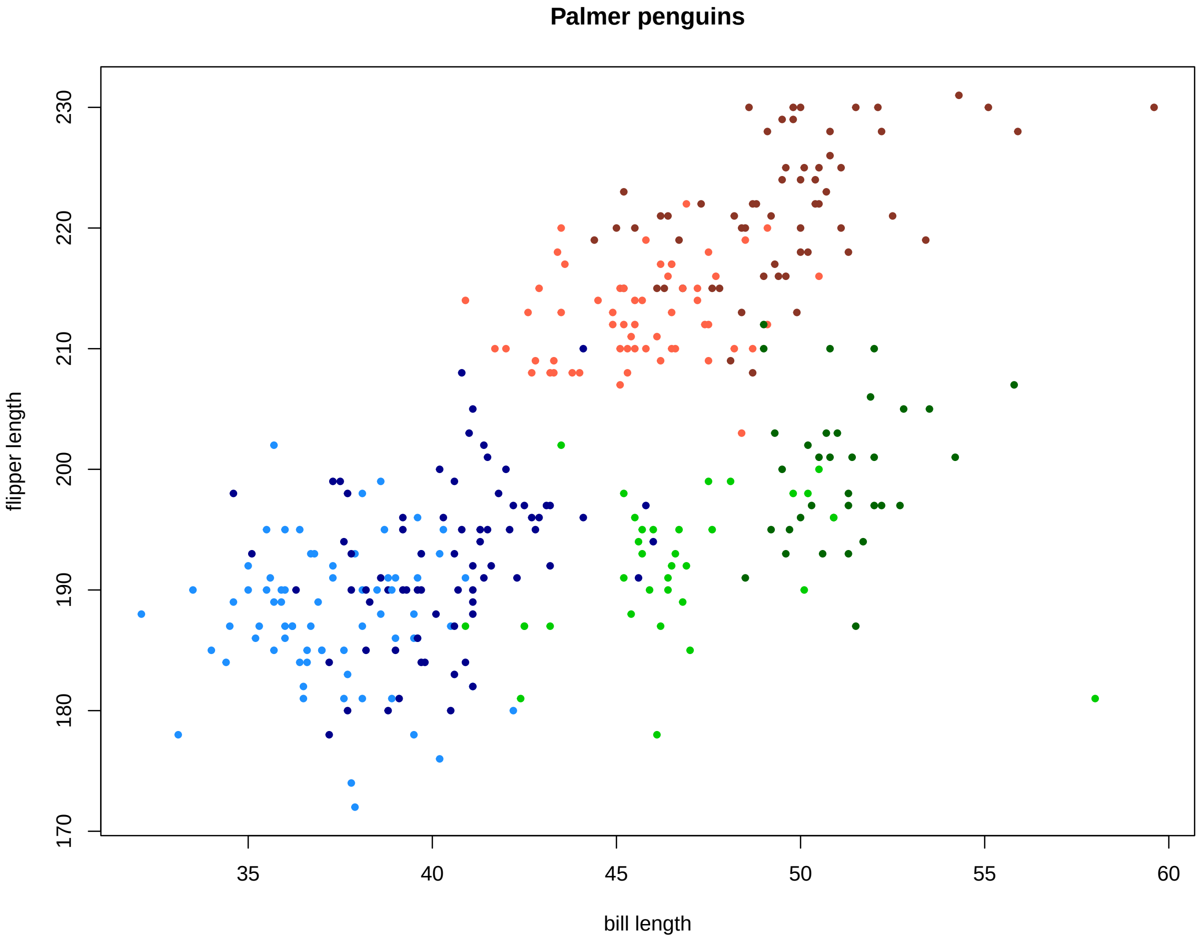
from the trendline between flipper length and body mass, we can say generally say that fish with longer flipper's tend to have a higher mass. we can also see that bill length and flipper length are positively correlated too, where penguins with longer bills generally have longer flippers. and from the distinction in sex we see that male penguins (blue dots) generally have higher mass and longer flippers than females.

d) Reproduce the plot in Figure 1 below. The colors used for the dots are `dodgerblue1` and `darkblue` for Adelie, `green3` and `darkgreen` for Chinstrap, and `tomato1` and `tomato4` for Gentoo.

```r
#same method as previous question, just with 5 if else statements for the species sex combinations.



penguins_colors <- with(penguins_f,
                        ifelse(species == "Adelie" & sex == "female", "dodgerblue1",
                        ifelse(species == "Adelie" & sex == "male", "darkblue",
                        ifelse(species == "Chinstrap" & sex == "female", "green3",
                        ifelse(species == "Chinstrap" & sex == "male", "darkgreen",
                        ifelse(species == "Gentoo" & sex == "female", "tomato1", "tomato4")))))
)
plot(penguins_f$bill_length_mm, penguins_f$flipper_length_mm,
     col = penguins_colors,
     pch = 20,
     xlab = "bill length",
     ylab = "flipper length",
     main = "Palmer penguins")
```



e) Finally, we want to assess whether the body weight measurements (`body_mass_g`) can reasonably be assumed to follow a normal distribution. To do this, use quantile-quantile (Q-Q) plots. Divide the plotting window into four panels, and create normal Q-Q plots for each of the three species (Adelie, Chinstrap, and Gentoo), and for the entire dataset Use the species name as the title for each plot. Also, include a reference line in each plot to help assess deviations from normality. After generating the plots, comment on your observations regarding the normality of body weight distributions across species and for the dataset as a whole.

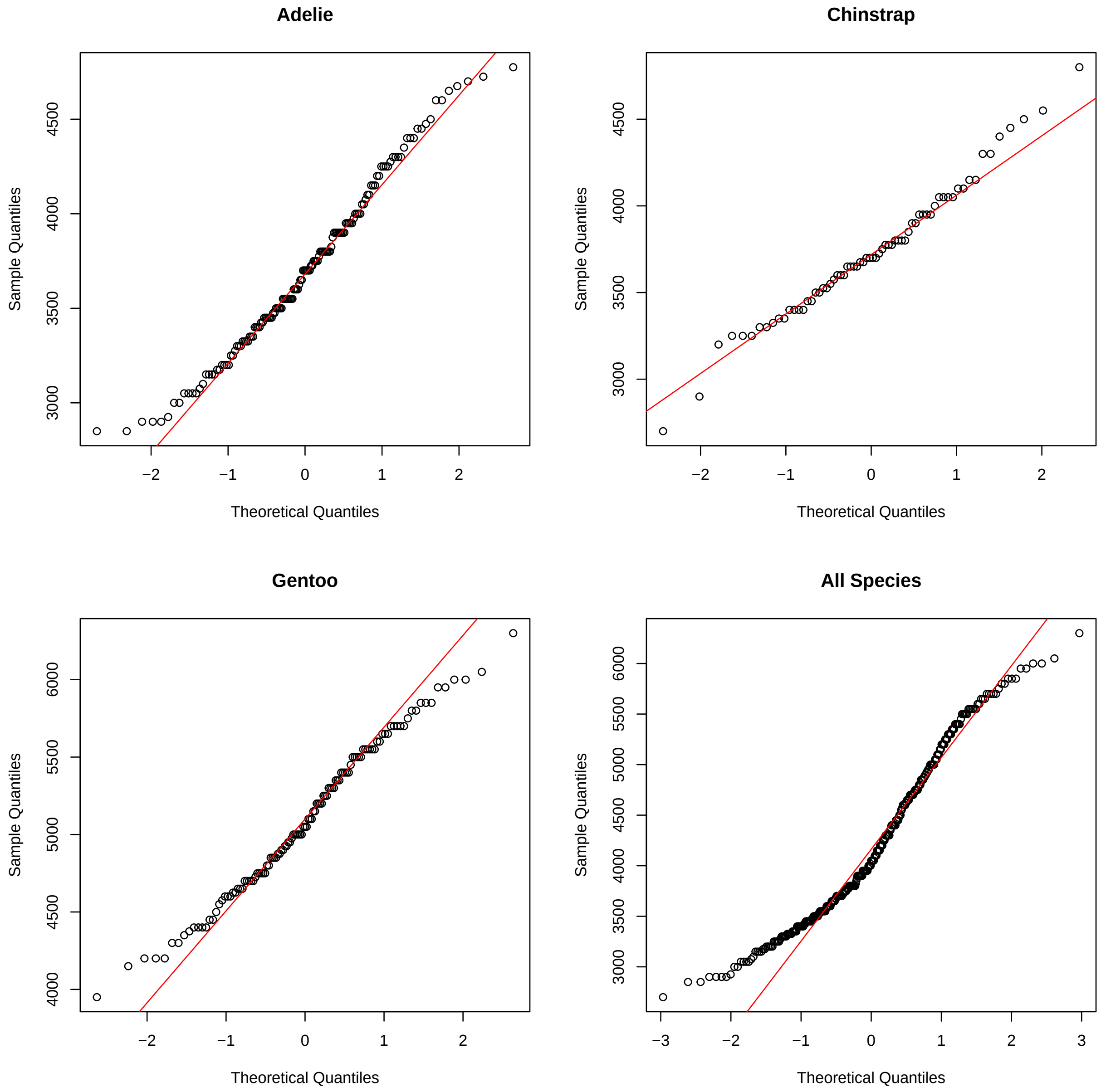
```
#divide into 4 subplots
par(mfrow = c(2,2))

qqnorm(penguins_f$body_mass_g[penguins_f$species == "Adelie"],
       main = "Adelie")
qqline(penguins_f$body_mass_g[penguins_f$species == "Adelie"],
       col = "red")

qqnorm(penguins_f$body_mass_g[penguins_f$species == "Chinstrap"],
       main = "Chinstrap")
qqline(penguins_f$body_mass_g[penguins_f$species == "Chinstrap"],
       col = "red")
```

```
qqnorm(penguins_f$body_mass_g[penguins_f$species == "Gentoo"],
       main = "Gentoo")
qqline(penguins_f$body_mass_g[penguins_f$species == "Gentoo"],
       col = "red")

qqnorm(penguins_f$body_mass_g,
       main = "All Species")
qqline(penguins_f$body_mass_g,
       col = "red")
```



Adelie: the QQ plot is pretty linear with heavier tails with the high points falling below the line, the distribution can be considered normal

Chinstrap: the QQ Plot is also pretty linear with tails pretty similar to the Adelie, but the high points rise above the line mirroring what happens to the Adelie; thus it can be considered a normal distribution.

Gentoo: the QQ- plot is linear but the qqline doesn't strictly follow the trend of data points, therefore it is not considered a normal distribution

All species: the points do not form a linear pattern, therefore the data could never fit the qq line; thus it is not a normal distribution

Question 2

The file `25Fhw3q2` has four simulated samples of size 30 obtained from the following distributions

- Standard logistic, (`rlogis(30)`)
- Exponential with default parameter, (`rexp(30)`)
- Uniform in (0,10), (`runif(30, min = 0, max = 10)`)
- Cauchy with default parameter (`rcauchy(30)`)

You have to identify which is which using quantile plots. Since you will need to draw quantile plots with respect to distributions other than the normal, it will be convenient to use a function named `qqPlot` in the package `car`. You will need to install this package. If you are using RStudio, select the `Packages` tab on the panel on the right and then select the `Install` tab. Type `car` on the pop-up window and click install. After installing, you need to load the package using `library(car)`.

The function `qqPlot` has syntax

```
qqPlot(x, dist = 'weibull', shape = 2)
```

for plotting a quantile graph of vector `x` with respect to the Weibull distribution with shape parameter 2. The default distribution for `qqPlot` is the normal distribution. You can find more details in the help for `qqPlot`. By default, this function draws confidence bands which I find in many cases of little use, and in some cases misleading. If you don't want them in your graph, add `envelope = FALSE` in your call.

Explain clearly the reasons for your choices.

```
  library(car)
```

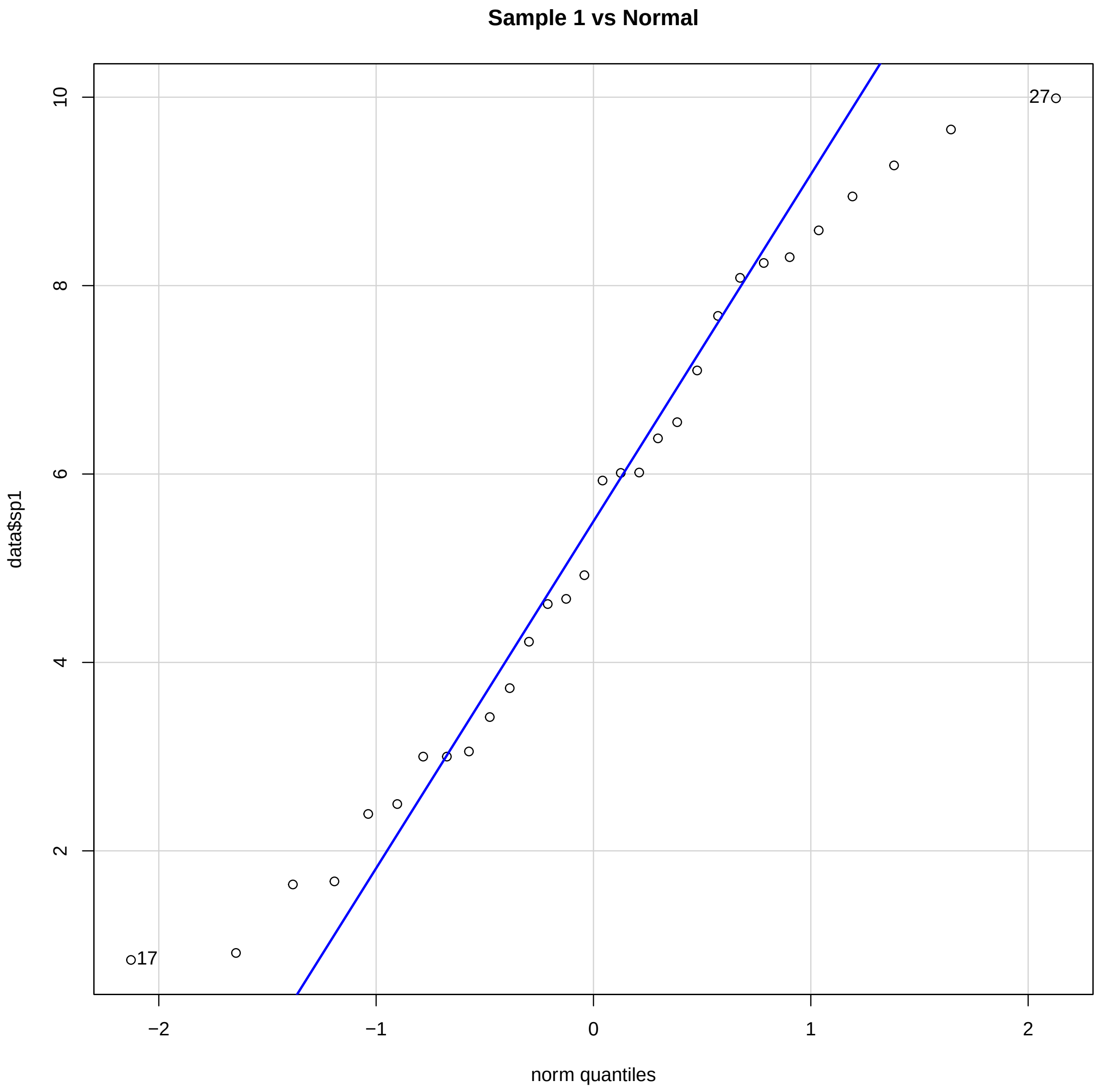
```
## Loading required package: carData
```

```
  #read the data in the file
  data <- read.table("25Fhw3q2")
  data
```

```
##          sp1         sp2        sp3           sp4
## 1  8.0822832 -1.87345572 1.47943763  -1.201827294
## 2  6.0115870 -0.10684920 1.02941205  -0.411370599
## 3  1.6758071  0.77640160 1.26492266  -0.933271774
## 4  9.2760035  1.22089989 0.58627673  -0.373836501
## 5  9.6571759 -0.27252832 0.27243804   7.268637283
## 6  5.9306926 -0.72699251 0.65470832   0.576115732
## 7  4.2198649 -2.10758804 0.19994573   1.632949883
## 8  2.4968691  1.43343114 0.12267648 -34.497961532
## 9  0.9162216  1.93468328 0.01854333  -1.150897900
## 10 6.5503488 -0.05009637 0.09688608  -0.032253538
## 11 8.5863651 -4.01444936 0.24239142   0.108924002
## 12 4.6745191  2.68957124 1.52000999   0.518953823
## 13 4.6196616 -3.95324932 0.50959529  -1.094246081
## 14 3.0001588  1.31533134 0.73325041   0.003570707
## 15 4.9260454 -0.78069595 0.67416558   2.810903107
## 16 3.4200334 -0.58490637 0.23948259   8.447741838
## 17 0.8418365 -1.54968229 0.37221023   2.727599720
## 18 7.0992754 -0.15436402 2.79396787   3.161457330
## 19 1.6431312 -1.36495115 0.58245838   0.630061381
## 20 7.6784329  0.39968494 4.62719404  -0.153046945
## 21 3.7270352  0.90482640 2.04480020   6.830590339
## 22 3.0004073  2.71060995 0.08543558   1.652200486
## 23 6.3785203 -0.49133735 0.87585204  -0.109080537
## 24 2.3916441  0.27014440 1.84000739  -1.310468658
## 25 8.9464268 -2.17046636 0.60802911  -2.048699606
## 26 8.3017136 -2.54997535 1.54930020   2.772917911
## 27 9.9885455  1.10717281 2.17543772  -1.389708786
## 28 6.0158339  1.82762732 0.30543898  -7.178617910
## 29 3.0548080 -0.01334288 1.45580425   0.862083781
## 30 8.2401579 -0.39328345 1.51144476  20.596754579
```

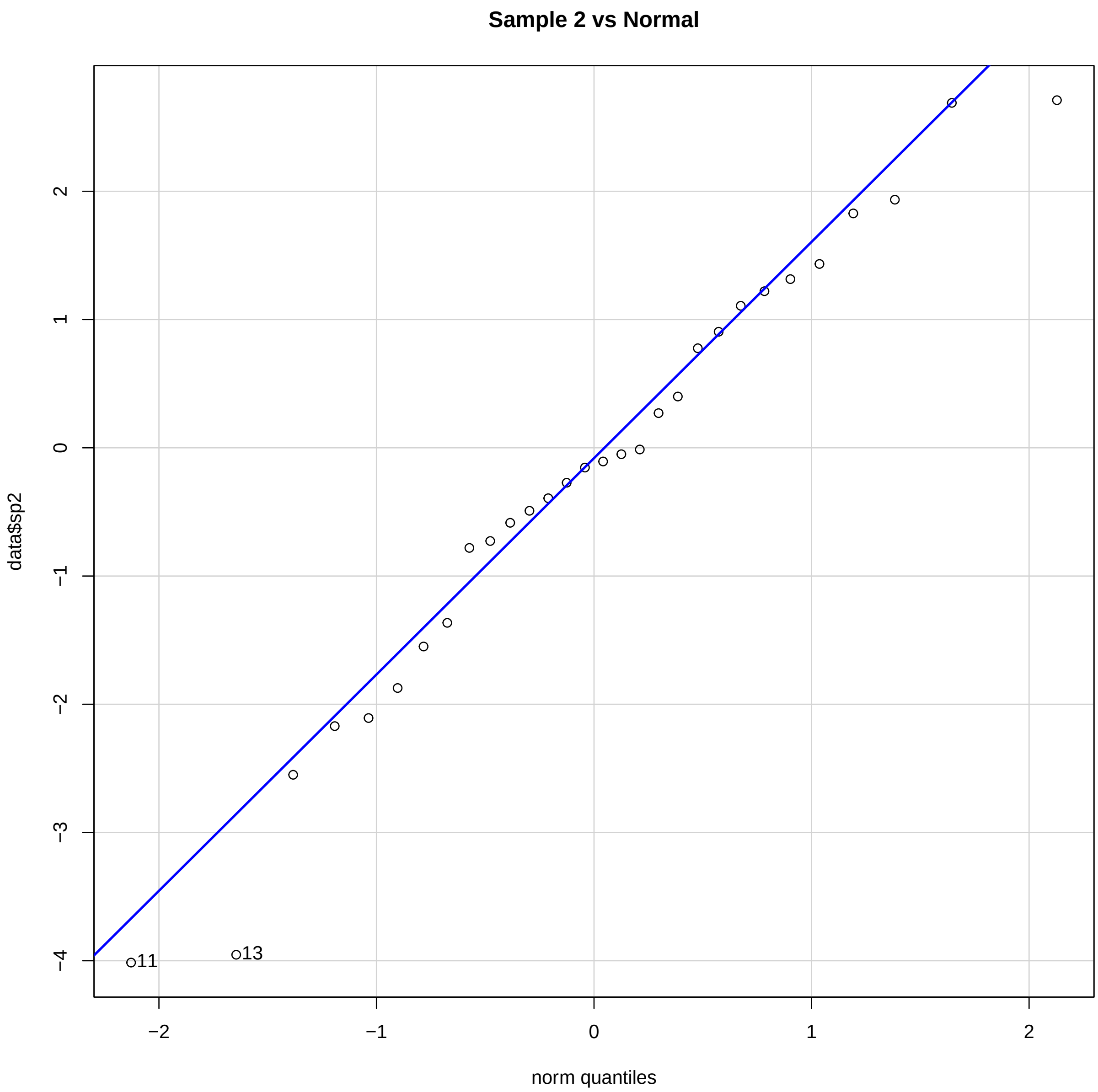
```
#par(mfrow = c(2,2))

qqPlot(data$sp1, dist = "norm", envelope = FALSE, main = "Sample 1 vs Normal")
```



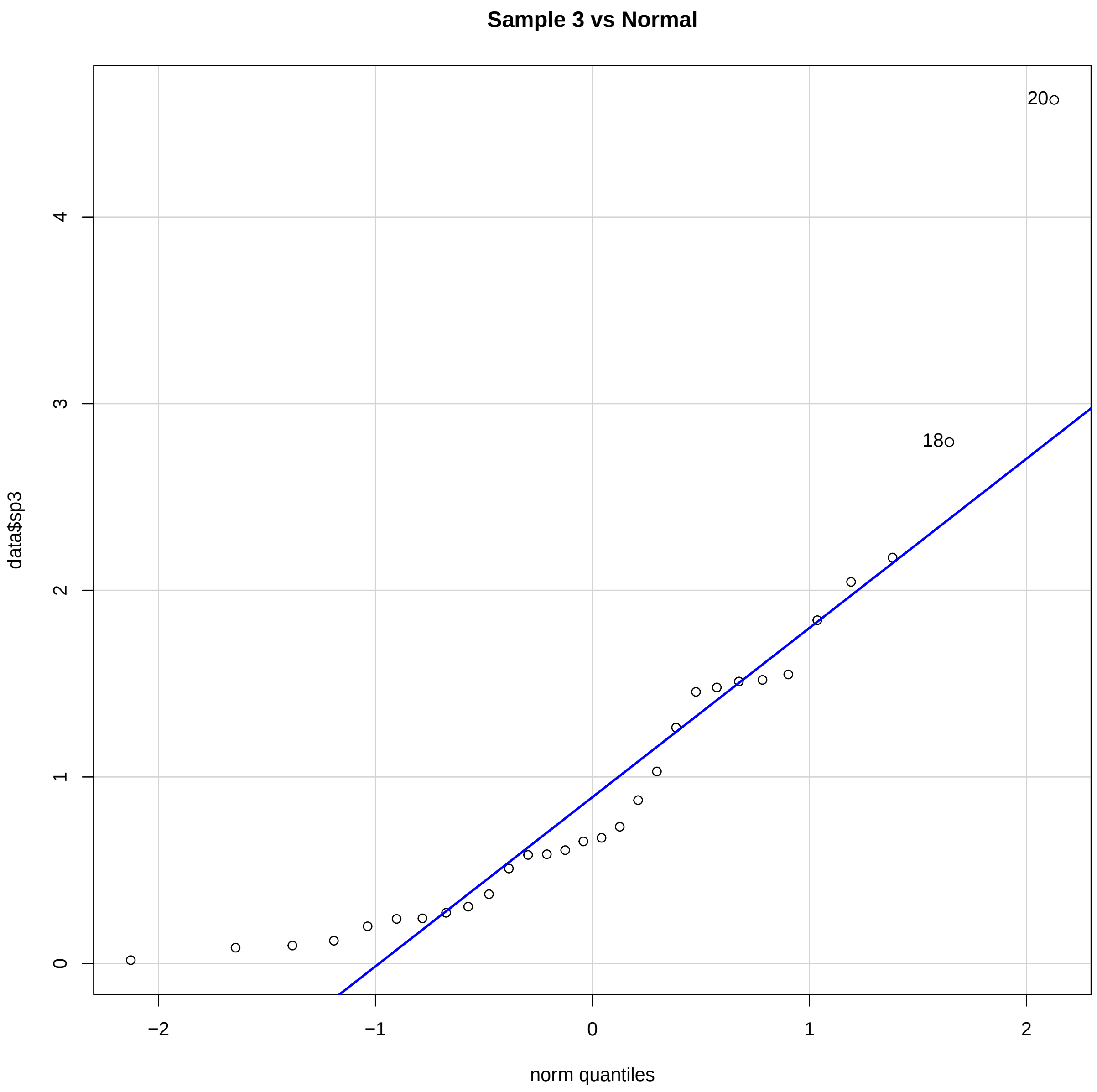
```
## [1] 27 17
```

```r
qqPlot(data$sp2, dist = "norm", envelope = FALSE, main = "Sample 2 vs Normal")
```



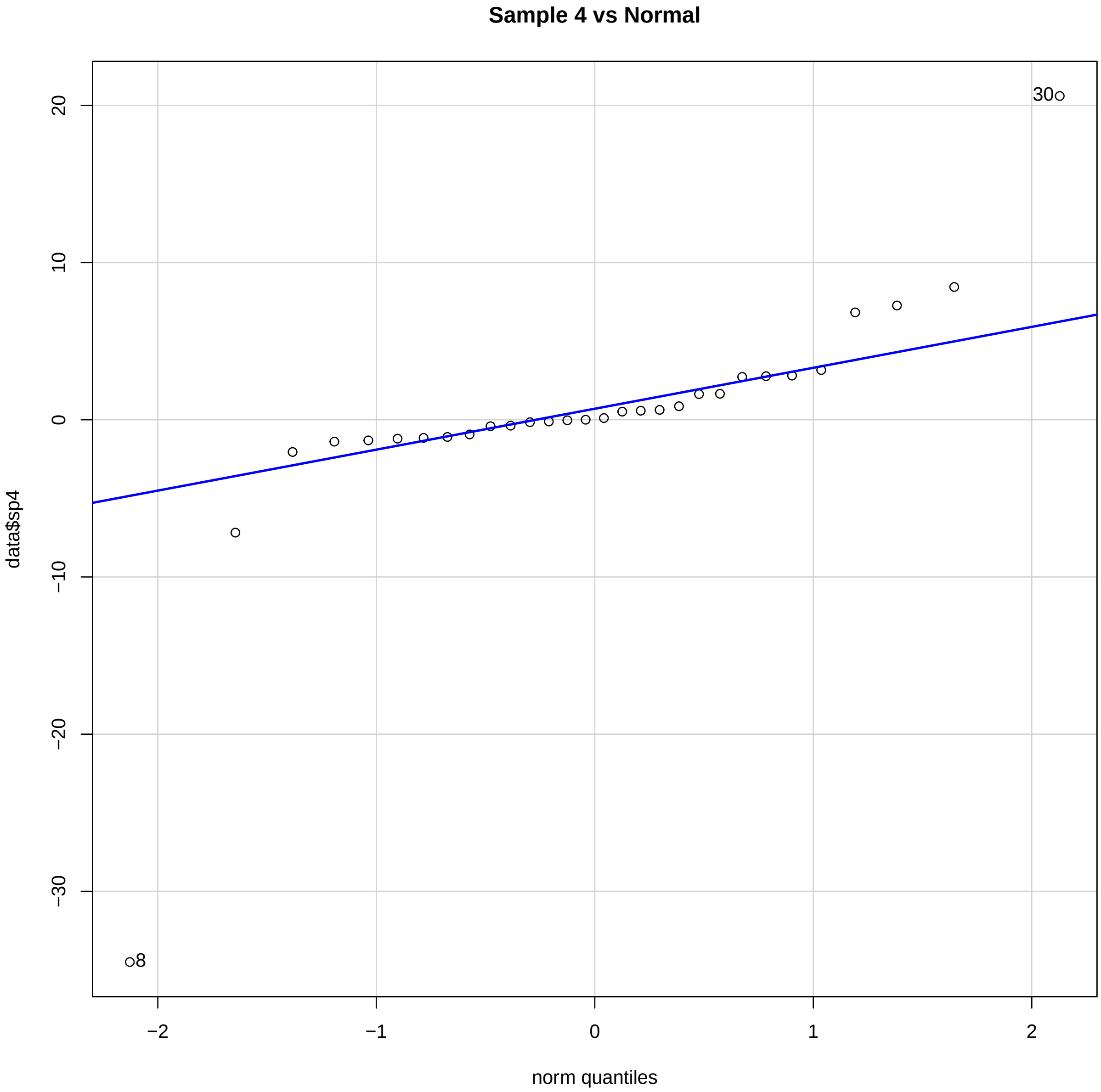
```
## [1] 11 13
```

```
qqPlot(data$sp3, dist = "norm", envelope = FALSE, main = "Sample 3 vs Normal")
```



```
## [1] 20 18
```

```
qqPlot(data$sp4, dist = "norm", envelope = FALSE, main = "Sample 4 vs Normal")
```



```
## [1]  8 30
```

I worked through this problem by pointing first checking the obvious plots. Sample 3 starts of near 0 and is never negative, and has a steadily increasing slope of points and its not linear at any point. this is a typical behavior of an exponential distribution

Sample 1 is the only one of the remaining 3 that doesn't have negative values, which follows the behavior of a uniform distribution with (0,10) range

sample 4 is very linear along the middle but has an extraordinary high rang eof values because of the extremely heavy outliers; even if most points have lower std. this is not expected of a logistic distribution therefore this is the cauchy distribution.

what remains is sample 2 being a logistic distribution, which seems reasonable as logistic distributions are mostly centered around 0, symmetric and have light tails.