Lab 4: Probability Distributions

DATA 606 - Statistics & Probability

Julia Ferris

2023-09-28

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

Exercise 1

Make a plot (or plots) to visualize the distributions of the amount of calories from fat of the options from these two restaurants. How do their centers, shapes, and spreads compare?

The distribution of the McDonald's data has a center around 250, it is right-skewed, and it has a wide spread. The distribution of the Dairy Queen data has a center somewhere between 220 and 260, is slightly more symmetrical than the McDonald's distribution, and has less spread than the McDonald's distribution. The center for the Dairy Queen distribution is not obvious from the graph, but the mean was about 260, and the median was 220.

```
data("fastfood", package='openintro')
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")
ggplot(mcdonalds, aes(x = cal_fat)) +
  geom_histogram(bins = 35)
ggplot(dairy_queen, aes(x = cal_fat)) +
  geom_histogram(bins = 35)

print("McDonald's")
```

```
## [1] "McDonald's"
```

```
mean(mcdonalds$cal_fat)
```

```
## [1] 285.614
```

```
median(mcdonalds$cal_fat)
```

```
## [1] 240
```

```r
sd(mcdonalds$cal_fat)
```

```
## [1] 220.8993
```

```r
var(mcdonalds$cal_fat)
```

```
## [1] 48796.49
```

```r
summary(mcdonalds$cal_fat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.0   160.0   240.0   285.6   320.0  1270.0
```

```r
print("Dairy Queen")
```

```
## [1] "Dairy Queen"
```

```r
mean(dairy_queen$cal_fat)
```

```
## [1] 260.4762
```

```r
median(dairy_queen$cal_fat)
```

```
## [1] 220
```

```r
sd(dairy_queen$cal_fat)
```

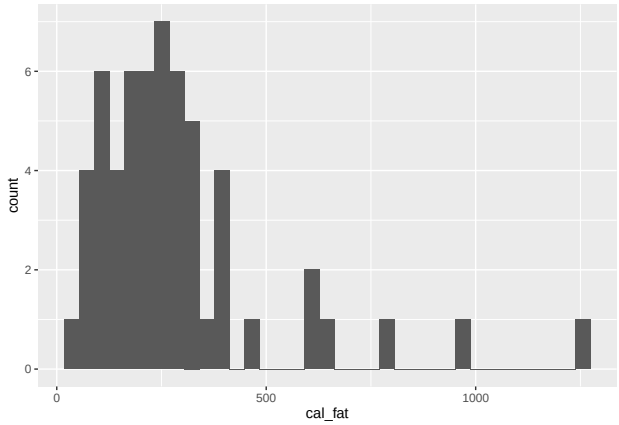
```
## [1] 156.4851
```

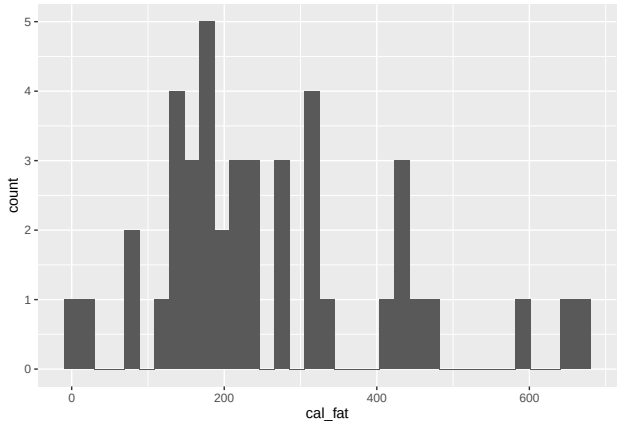
```r
var(dairy_queen$cal_fat)
```

```
## [1] 24487.57
```

```r
summary(dairy_queen$cal_fat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   160.0   220.0   260.5   310.0   670.0
```





Exercise 2

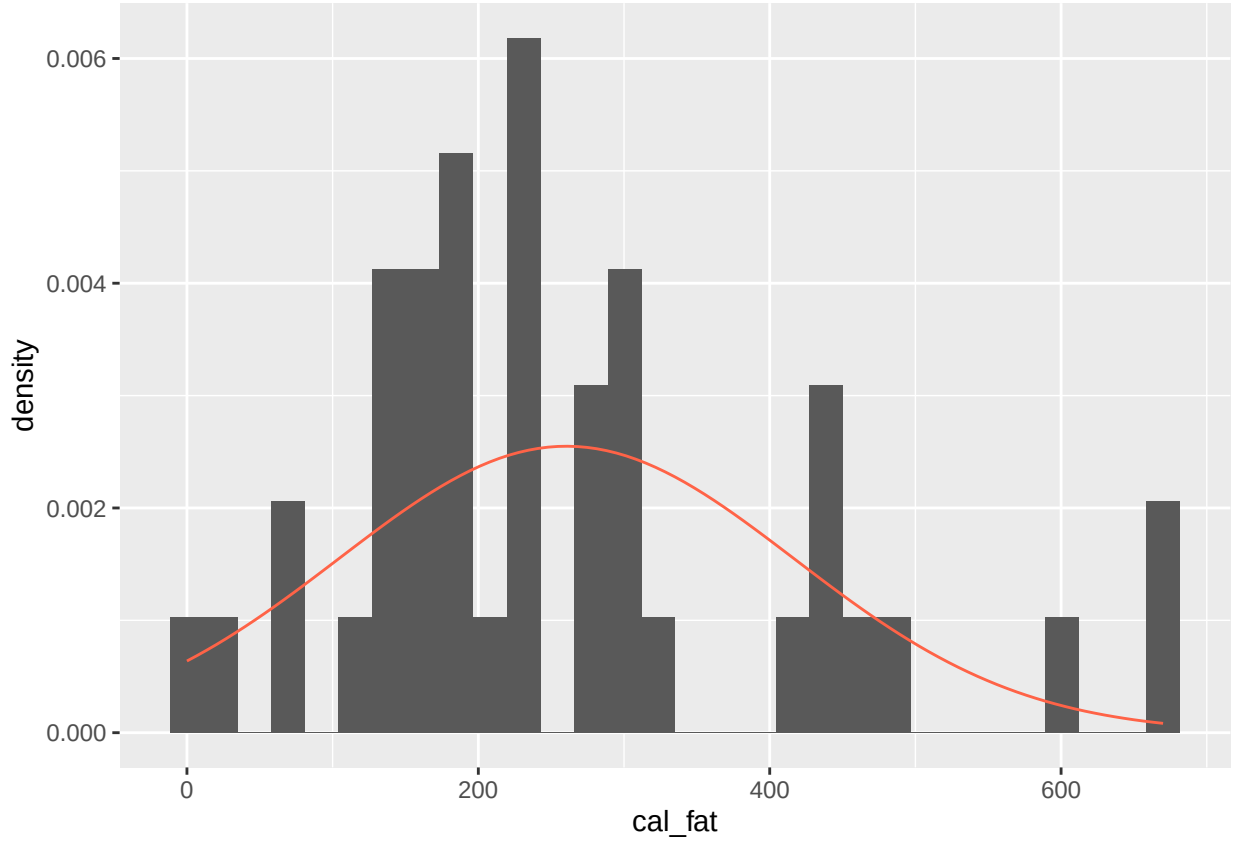
Based on the this plot, does it appear that the data follow a nearly normal distribution?

The data does not appear to follow a normal distribution. The shape of the curve is similar to the shape of the bars, but the height of the curve is much lower than the height of the bars. I would not claim that this is nearly a normal distribution.

```
dqmean <- mean(dairy_queen$cal_fat)
dqsd   <- sd(dairy_queen$cal_fat)
ggplot(data = dairy_queen, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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



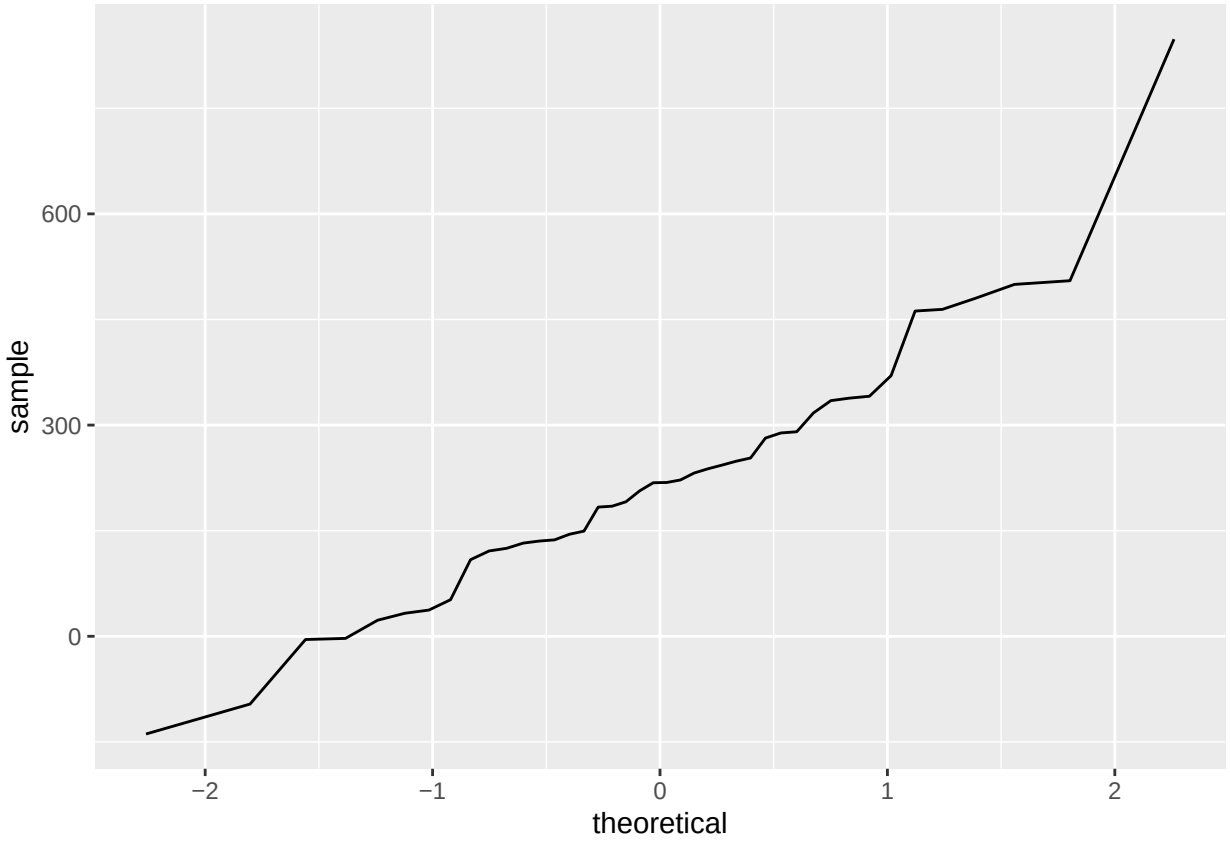
Exercise 3

Make a normal probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data? (Since sim_norm is not a data frame, it can be put directly into the sample argument and the data argument can be dropped.)

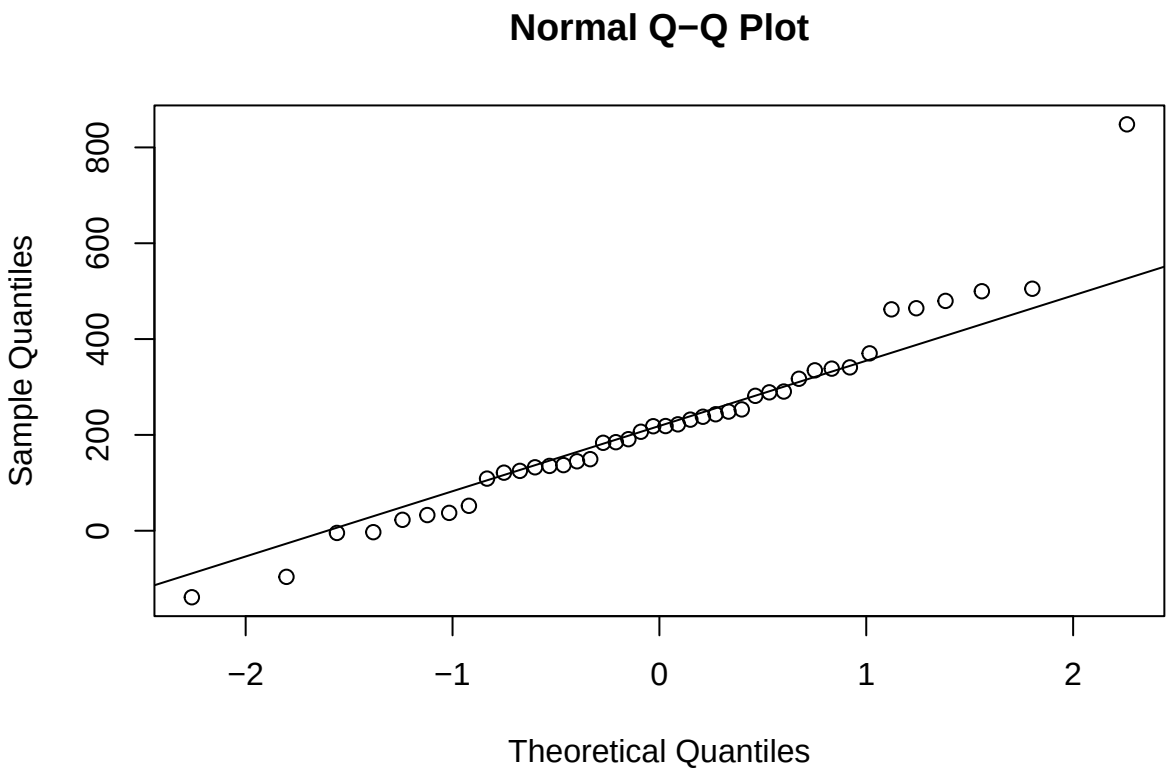
Not all of the points fall on the line as seen in the first Q-Q plot. Most of the middle values fall on the line, but the ones near both ends seem to stray.

This plot is more normally distributed than the plot for the real data. The second Q-Q plot below shows many values far from the line on the right side of the plot, indicating it is not normally distributed.

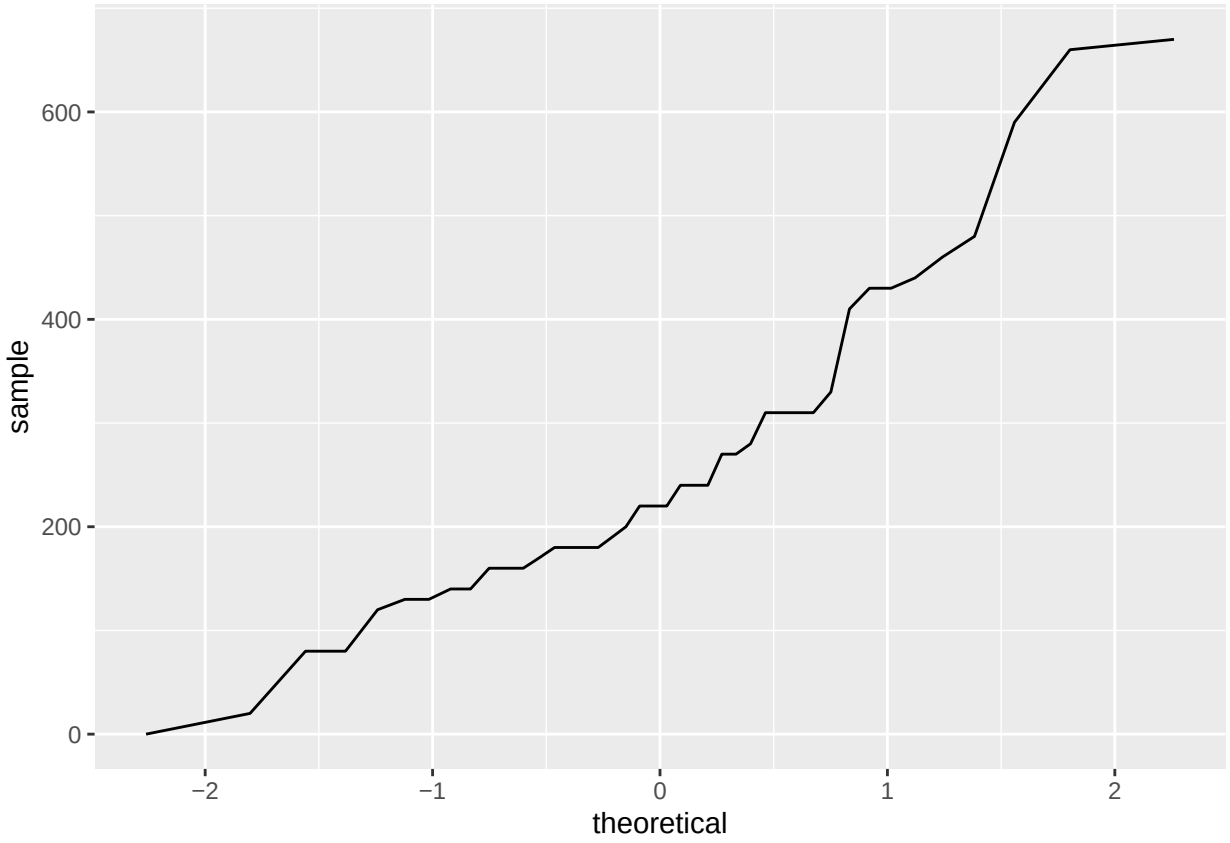
```
sim_norm <- rnorm(n = nrow(dairy_queen), mean = dqmean, sd = dqsd)
ggplot(mapping = aes(sample = sim_norm)) +
  geom_line(stat = "qq")
```



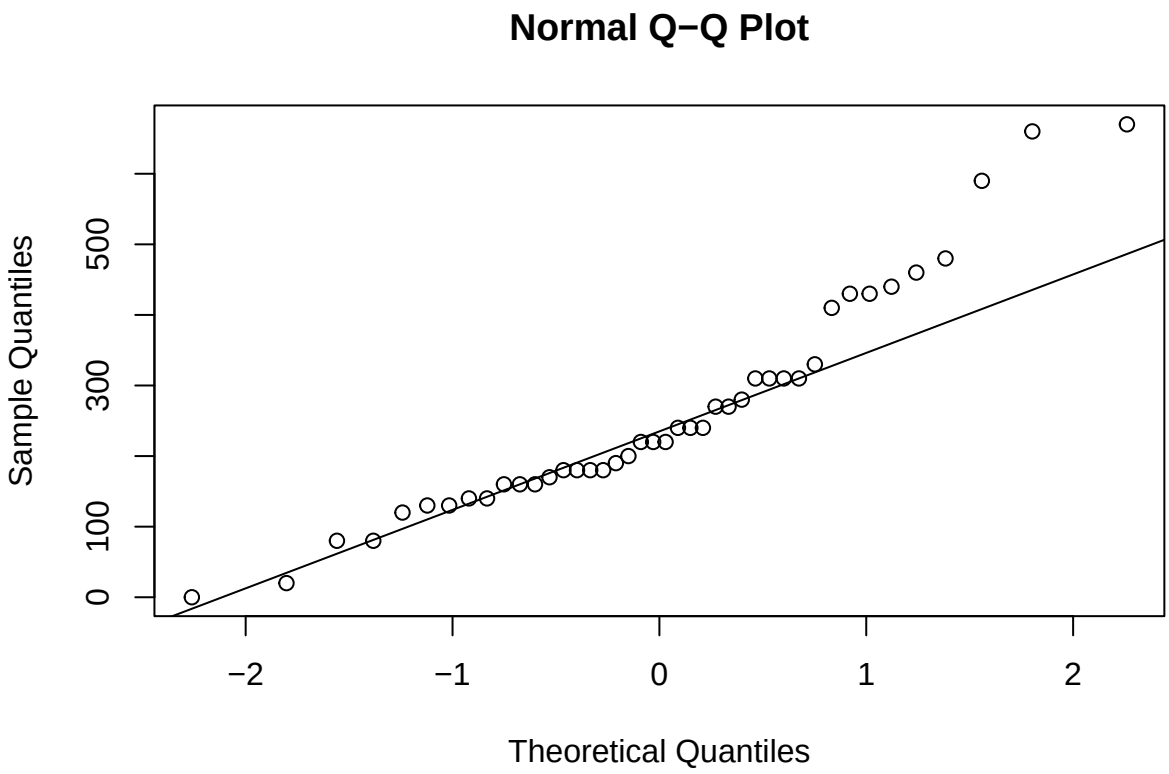

```
qqnorm(sim_norm)
qqline(sim_norm)
```




```
ggplot(data = dairy_queen, aes(sample = cal_fat)) +
  geom_line(stat = "qq")
```




```
qqnorm(dairy_queen$cal_fat)
qqline(dairy_queen$cal_fat)
```

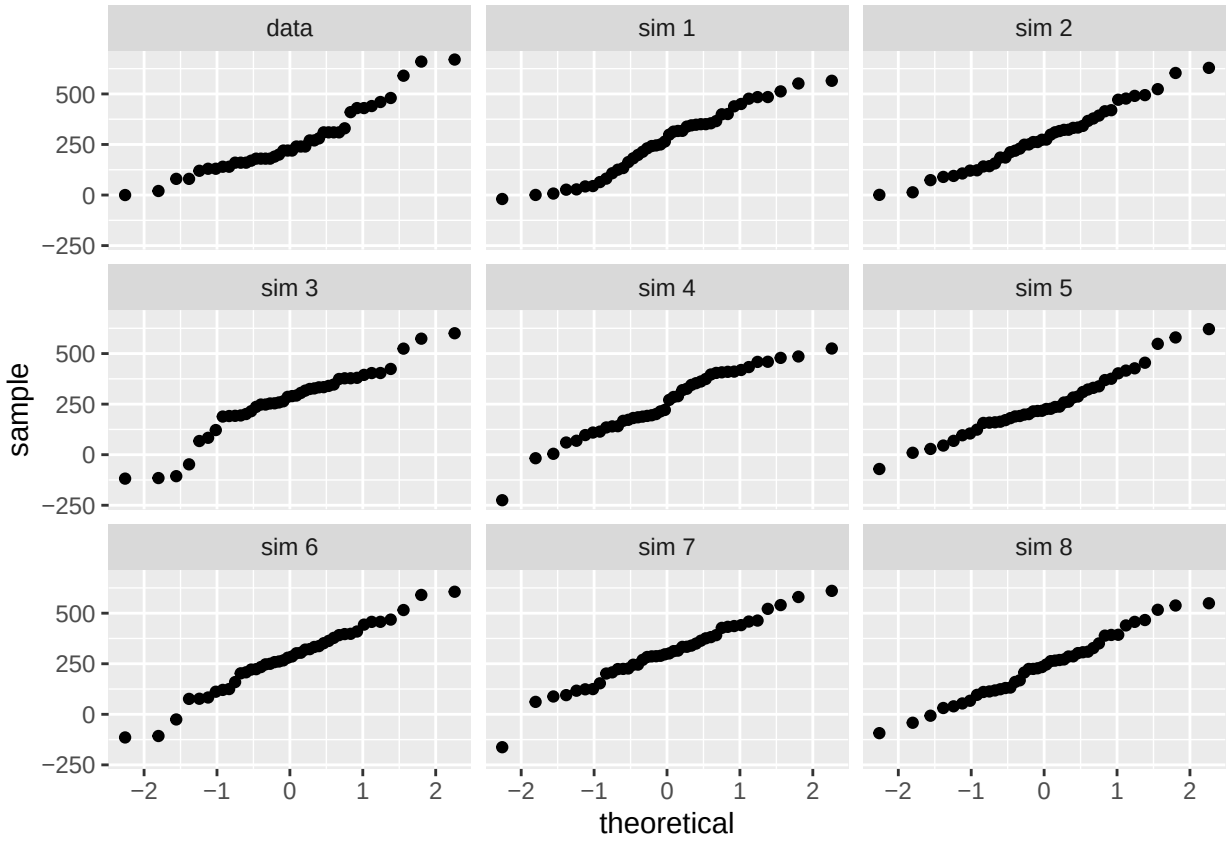


Exercise 4

Does the normal probability plot for the calories from fat look similar to the plots created for the simulated data? That is, do the plots provide evidence that the calories are nearly normal?

Yes, it looks pretty similar to the plots created for the simulated data. I would say that this does provide evidence that the calories are nearly normal because the original data has a graph very similar to all the other plot, and most of the data points in each of the graphs fall along the line. It is not exactly normal, but it is pretty close.

```
openintro::qqnormsim(sample = cal_fat, data = dairy_queen)
```

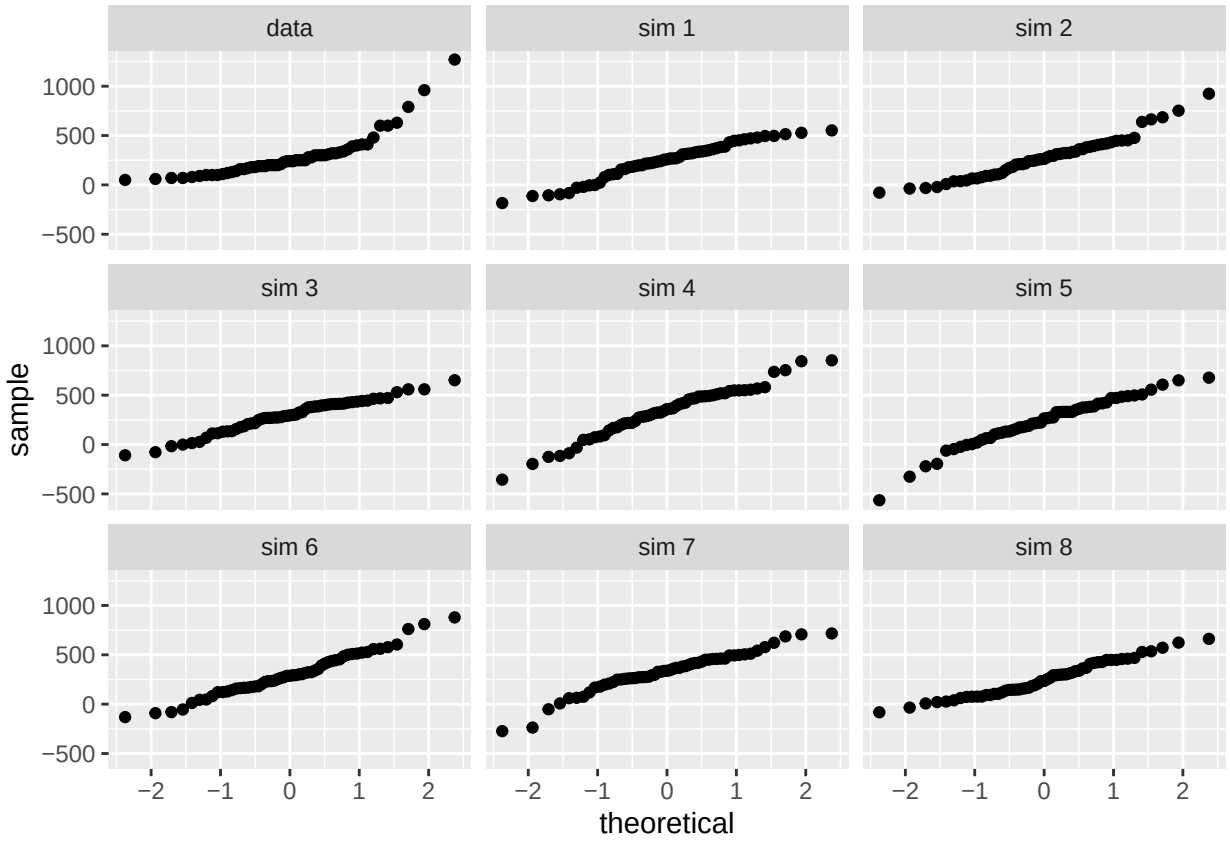


Exercise 5

Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

The data from McDonald's does not resemble the other graphs. I would say it is not normally distributed because of the differences and because many of the data points in the original graph do not fall on the line.

```
openintro::qqnormsim(sample = cal_fat, data = mcdonalds)
```



Exercise 6

Write out two probability questions that you would like to answer about any of the restaurants in this dataset. Calculate those probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which one had a closer agreement between the two methods?

1. If you were to randomly order an item from the Chick Fil-A menu, what is the probability of choosing an item with a cholesterol greater than 100 mg?

   Theoretical normal distribution probability: 0.3291412 Empirical distribution probability: 0.1486697 Difference (absolute value): 0.2180301

2. If you were to randomly order an item from Subway, what is the probability of choosing an item that contained total carbs less than 20 g?

   Theoretical normal distribution probability: 0.1111111 Empirical distribution probability: 0.2291667 Difference (absolute value): 0.08049694

The second question had more similar results than the first question. This would typically indicate that the data about total carbs for Subway was closer to a normal distribution than the data about cholesterol for Chick Fil-A. However, the Q-Q plots show that the chick Fil-A data about cholesterol had an outlier that affected the normality of the data, and the Q-Q plot for Subway was not normal.

```
chickFilA <- fastfood |> filter(restaurant == "Chick Fil-A")
chmean <- mean(chickFilA$cholesterol)
```

```r
chsd <- sd(chickFilA$cholesterol)
(chProb <- 1 - pnorm(q = 100, mean = chmean, sd = chsd))
```

```
## [1] 0.3291412
```

```r
(chActualProb <- chickFilA |>
  filter(cholesterol > 100) |>
  summarise(percent = n() / nrow(chickFilA)))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.111
```

```r
subway <- fastfood |> filter(restaurant == "Subway")
smean <- mean(subway$total_carb)
ssd <- sd(subway$total_carb)
(subProb <- pnorm(q = 20, mean = smean, sd = ssd))
```

```
## [1] 0.1486697
```

```r
(subActualProb <- subway |>
  filter(total_carb < 20) |>
  summarise(percent = n() / nrow(subway)))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.229
```

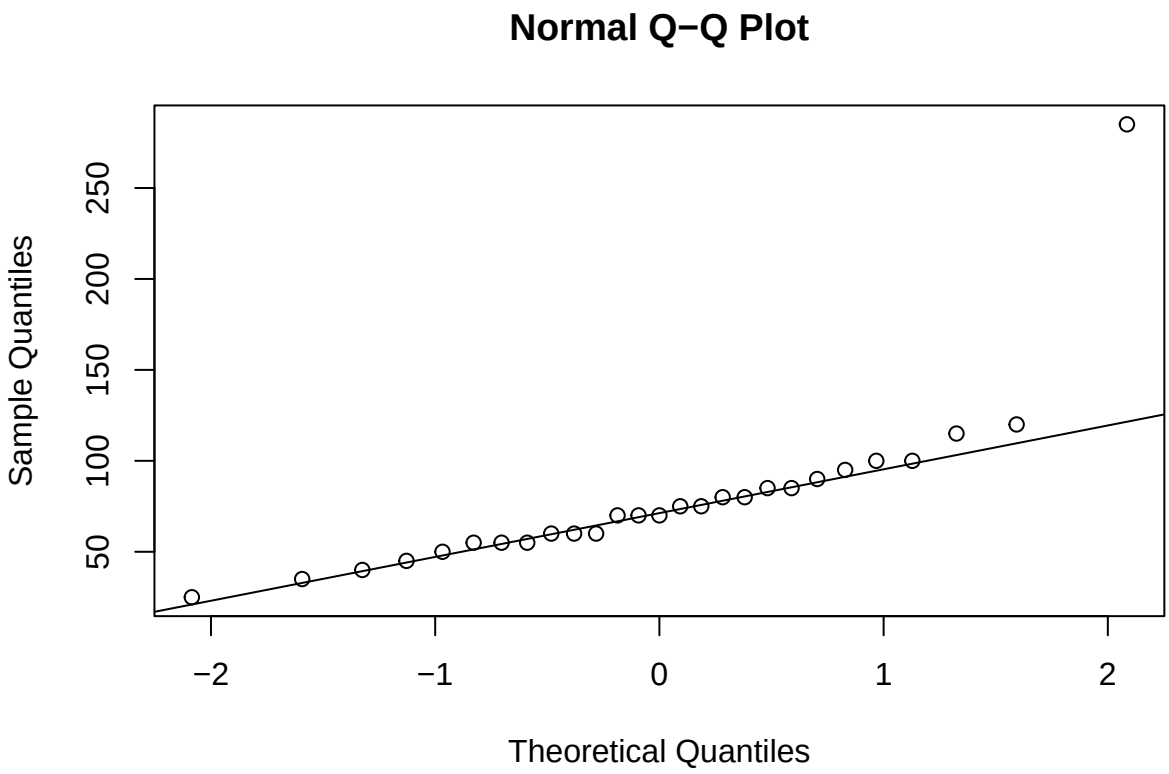
```r
chProb - chActualProb
```

```
##     percent
## 1 0.2180301
```

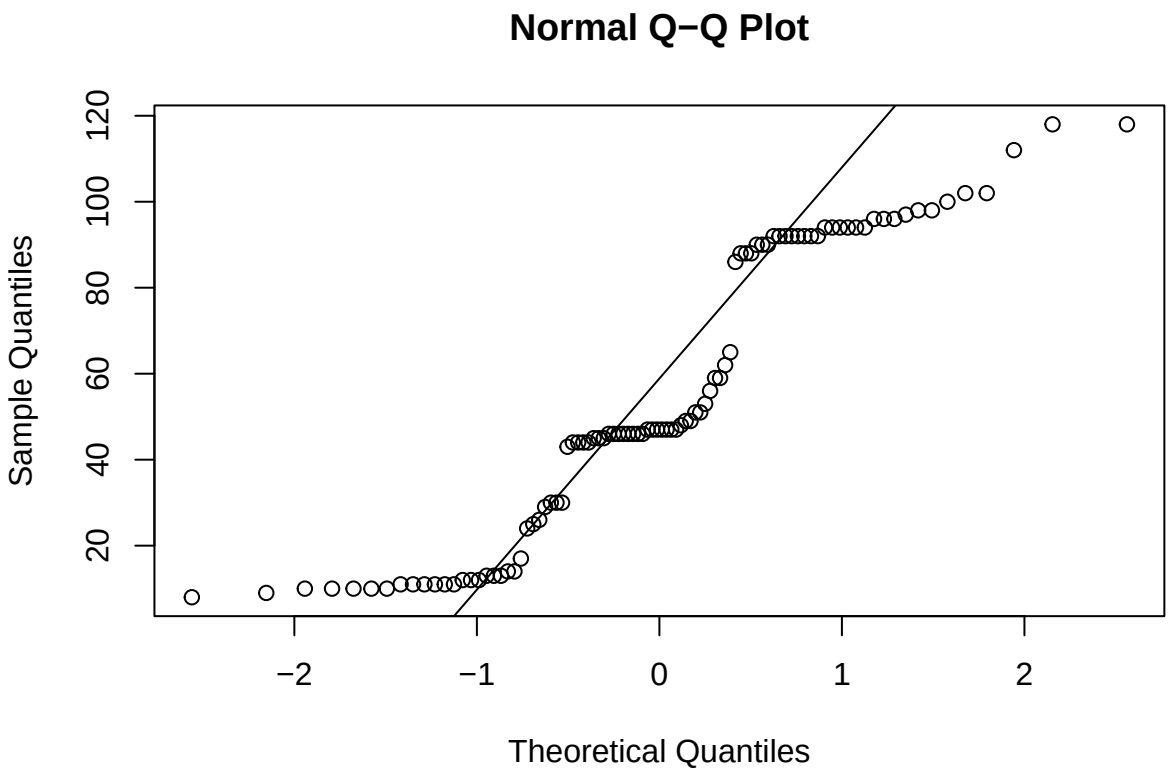
```r
subProb - subActualProb
```

```
##       percent
## 1 -0.08049694
```

```r
qqnorm(chickFilA$cholesterol)
qqline(chickFilA$cholesterol)
```




```
qqnorm(subway$total_carb)
qqline(subway$total_carb)
```



Exercise 7

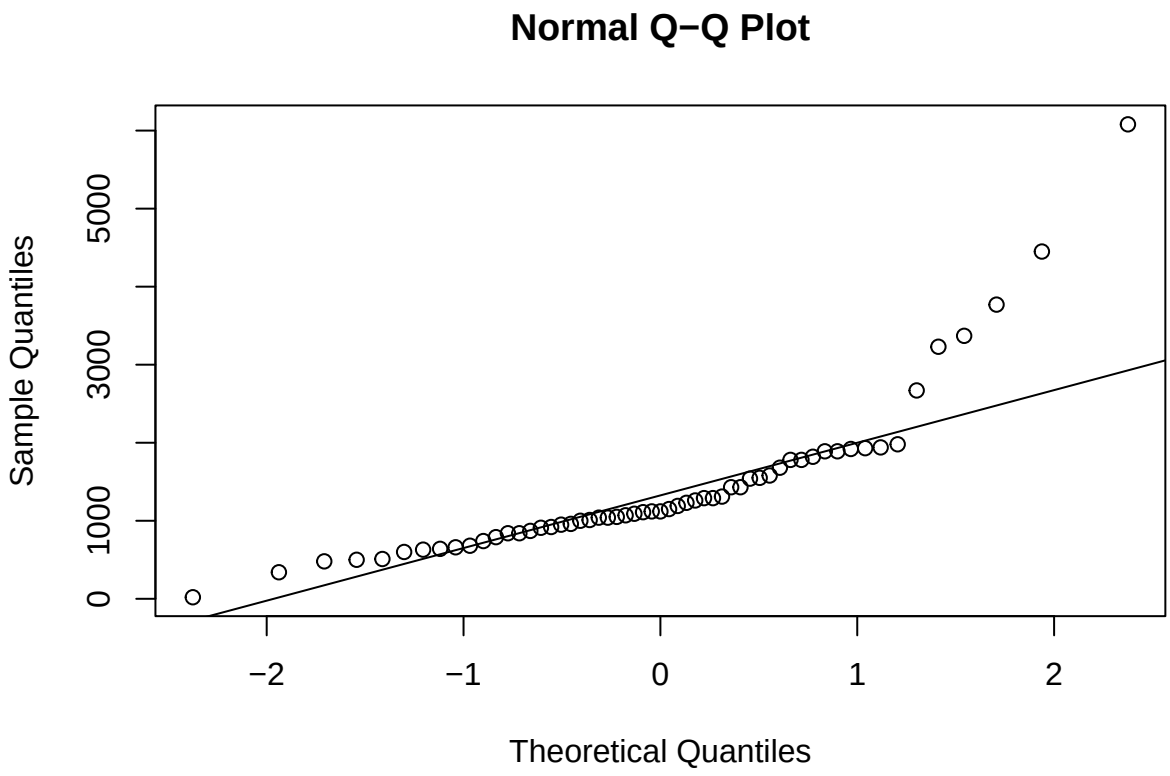
Now let's consider some of the other variables in the dataset. Out of all the different restaurants, which ones' distribution is the closest to normal for sodium?

Arby's had the closest to normal distribution for sodium. I verified my assumption with the Shapiro-Wilk Test. The p-value for Arby's was greater than 0.05, just like Burger King. Since it had a greater p-value than Arby's I decided it was the most normal.

```
unique(fastfood$restaurant)
```

```
## [1] "Mcdonalds"   "Chick Fil-A" "Sonic"       "Arbys"       "Burger King"
## [6] "Dairy Queen" "Subway"      "Taco Bell"
```

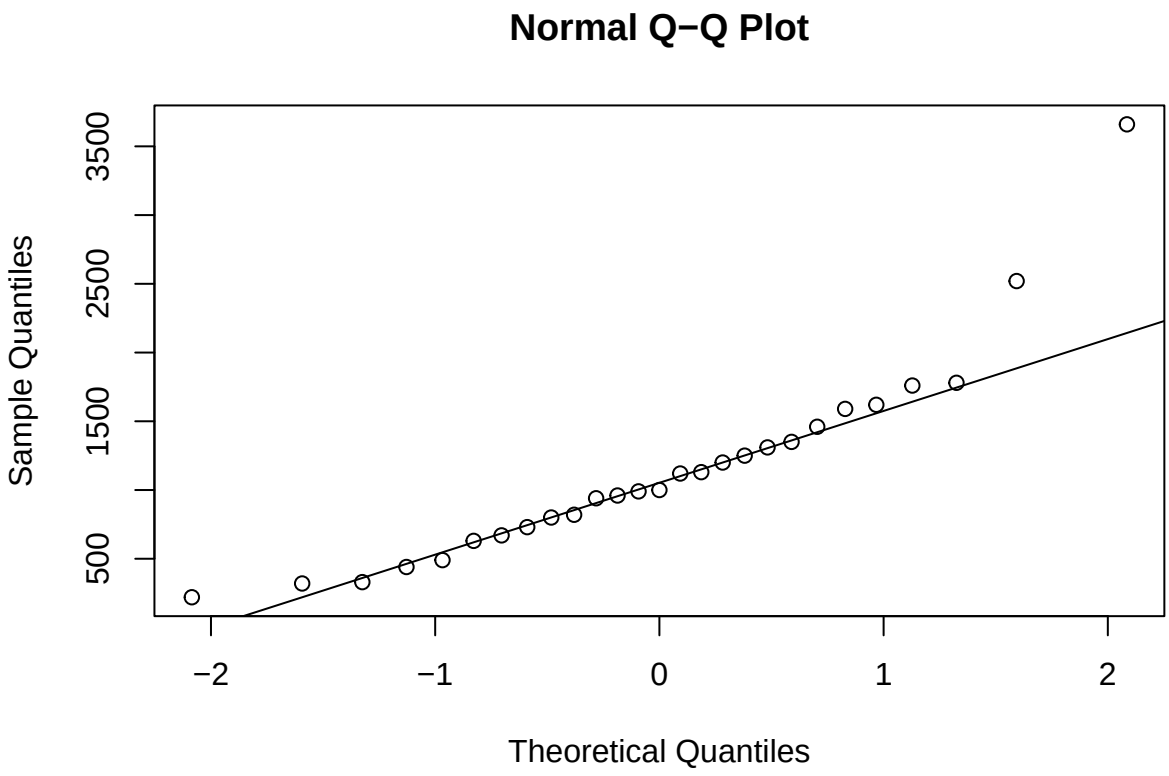
```
sonic <- fastfood |> filter(restaurant == "Sonic")
arbys <-fastfood |> filter(restaurant == "Arbys")
bk <- fastfood |> filter(restaurant == "Burger King")
tacoBell <- fastfood |> filter(restaurant == "Taco Bell")

qqnorm(mcdonalds$sodium)
qqline(mcdonalds$sodium)
```




```
shapiro.test(mcdonalds$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mcdonalds$sodium
## W = 0.76922, p-value = 4.458e-08
```

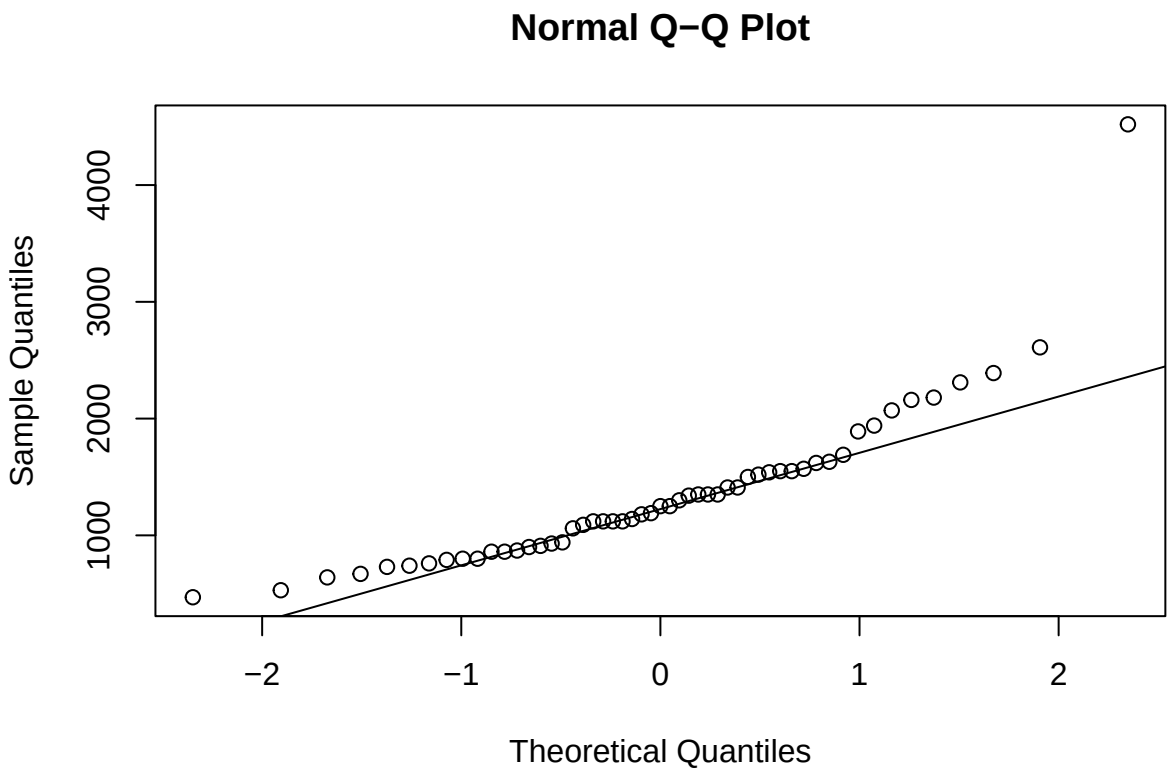
```
qqnorm(chickFilA$sodium)
qqline(chickFilA$sodium)
```




```
shapiro.test(chickFilA$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chickFilA$sodium
## W = 0.86663, p-value = 0.002503
```

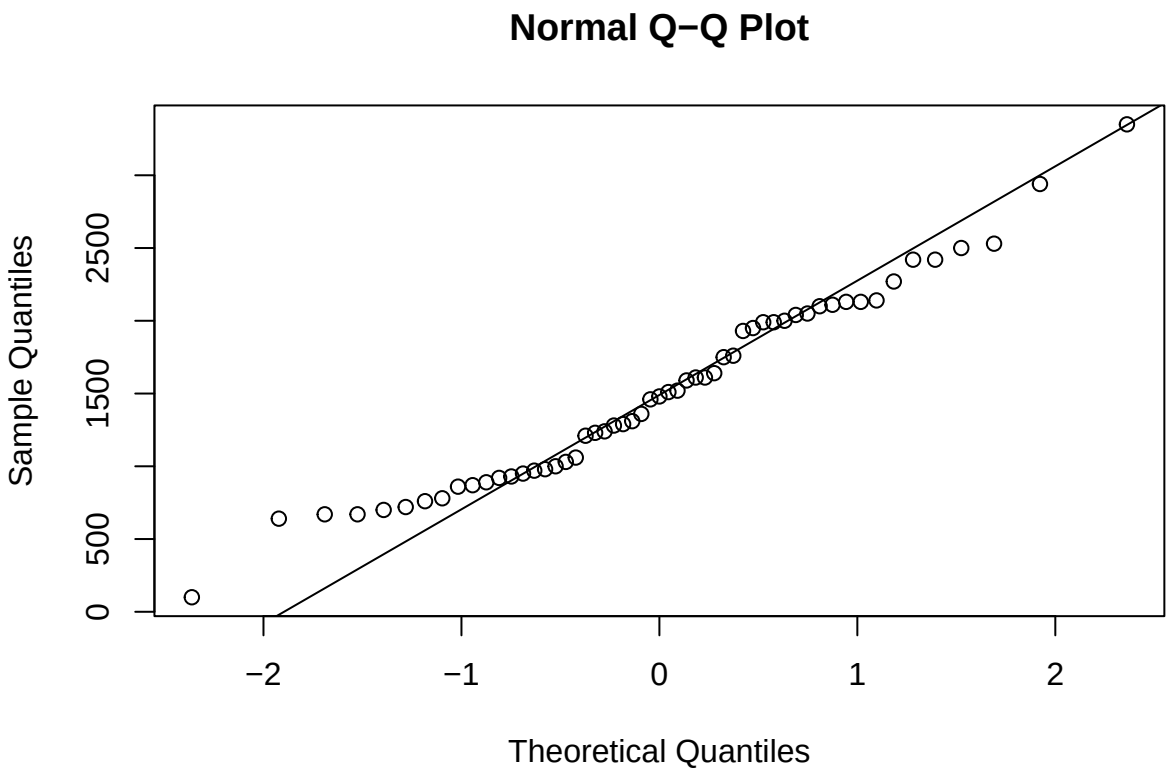
```
qqnorm(sonic$sodium)
qqline(sonic$sodium)
```




```
shapiro.test(sonic$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sonic$sodium
## W = 0.82286, p-value = 1.784e-06
```

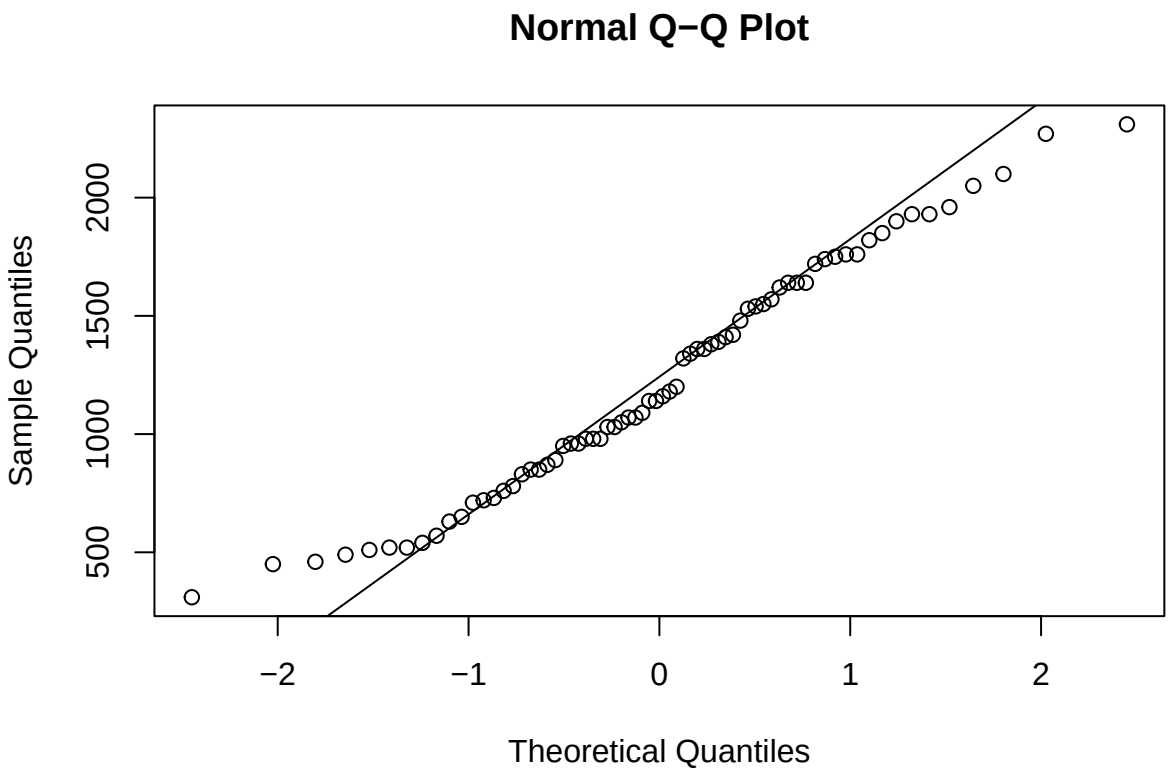
```
qqnorm(arbys$sodium)
qqline(arbys$sodium)
```




```
shapiro.test(arbys$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  arbys$sodium
## W = 0.97073, p-value = 0.1985
```

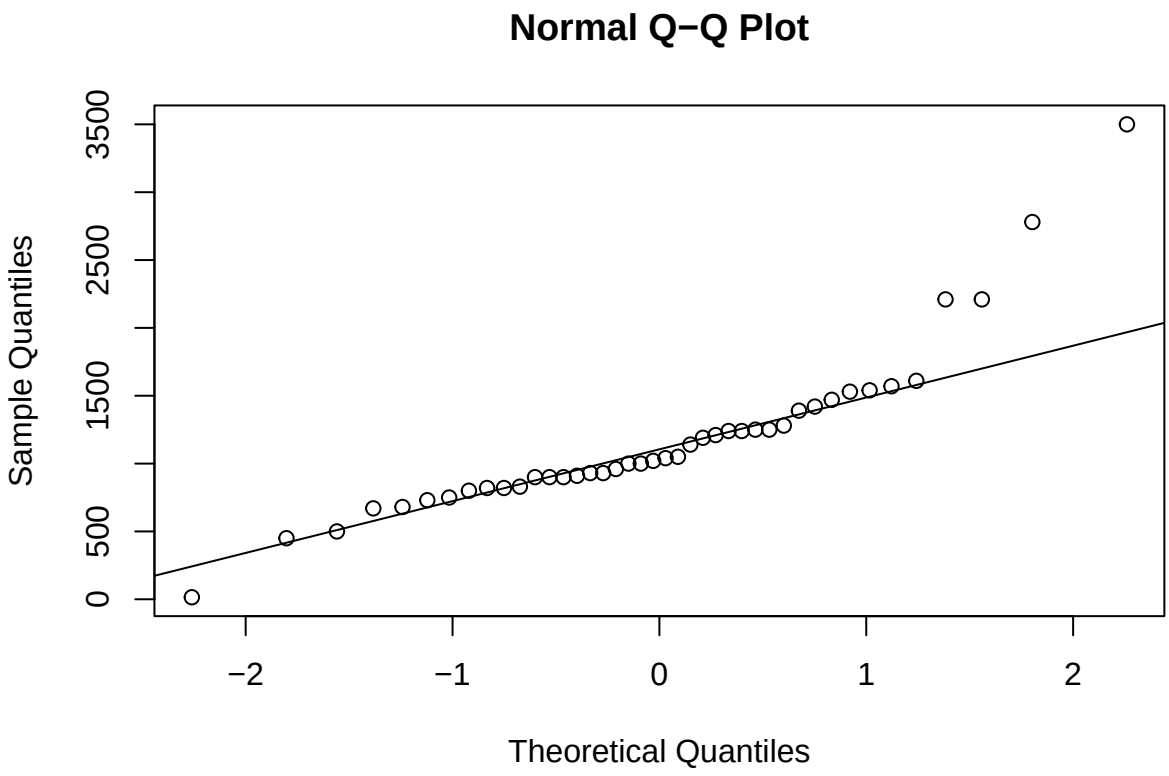
```
qqnorm(bk$sodium)
qqline(bk$sodium)
```




```
shapiro.test(bk$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bk$sodium
## W = 0.97291, p-value = 0.1331
```

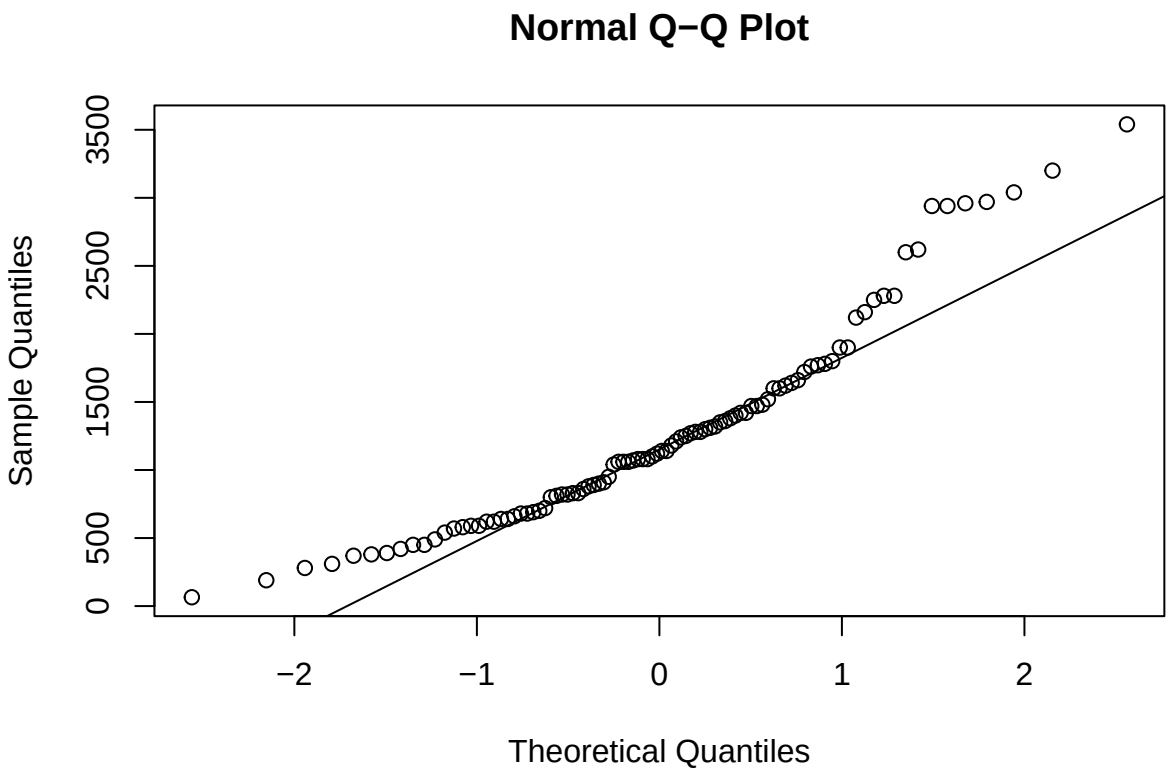
```
qqnorm(dairy_queen$sodium)
qqline(dairy_queen$sodium)
```




```
shapiro.test(dairy_queen$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dairy_queen$sodium
## W = 0.84504, p-value = 4.715e-05
```

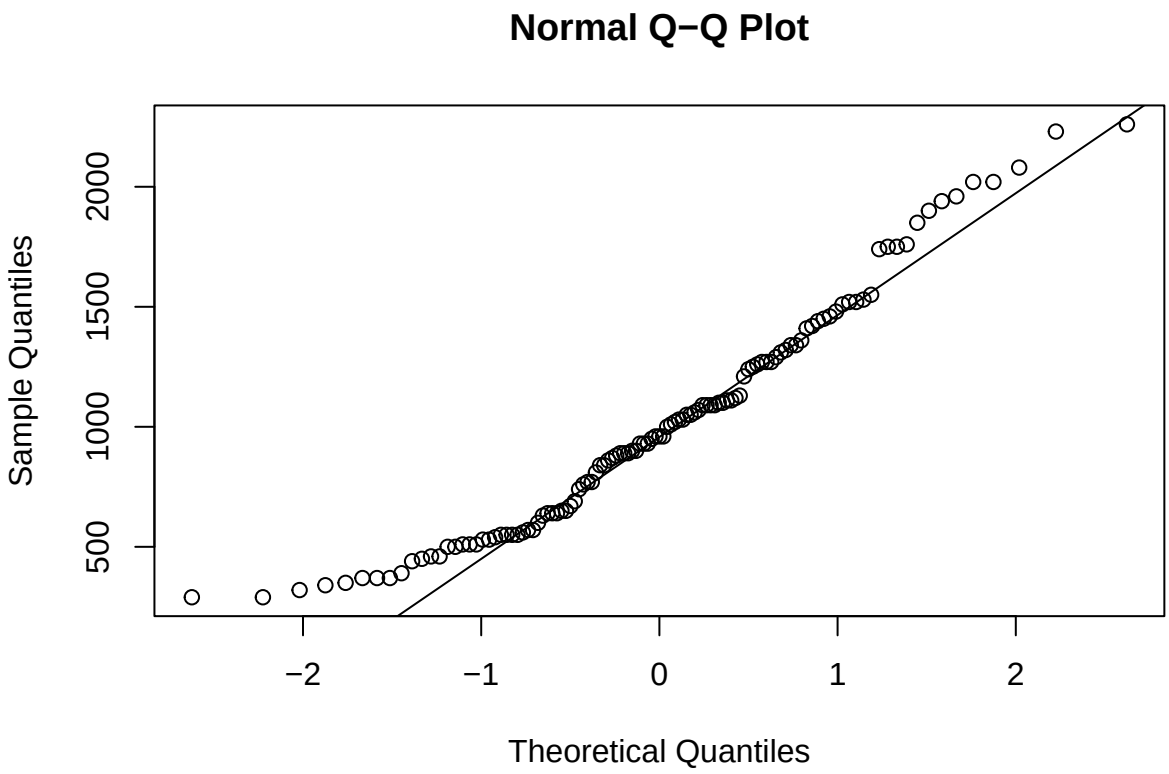
```
qqnorm(subway$sodium)
qqline(subway$sodium)
```




```
shapiro.test(subway$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subway$sodium
## W = 0.92175, p-value = 2.515e-05
```

```
qqnorm(tacoBell$sodium)
qqline(tacoBell$sodium)
```




```
shapiro.test(tacoBell$sodium)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tacoBell$sodium
## W = 0.95501, p-value = 0.000699
```

Exercise 8

Note that some of the normal probability plots for sodium distributions seem to have a step-wise pattern. Why do you think this might be the case?

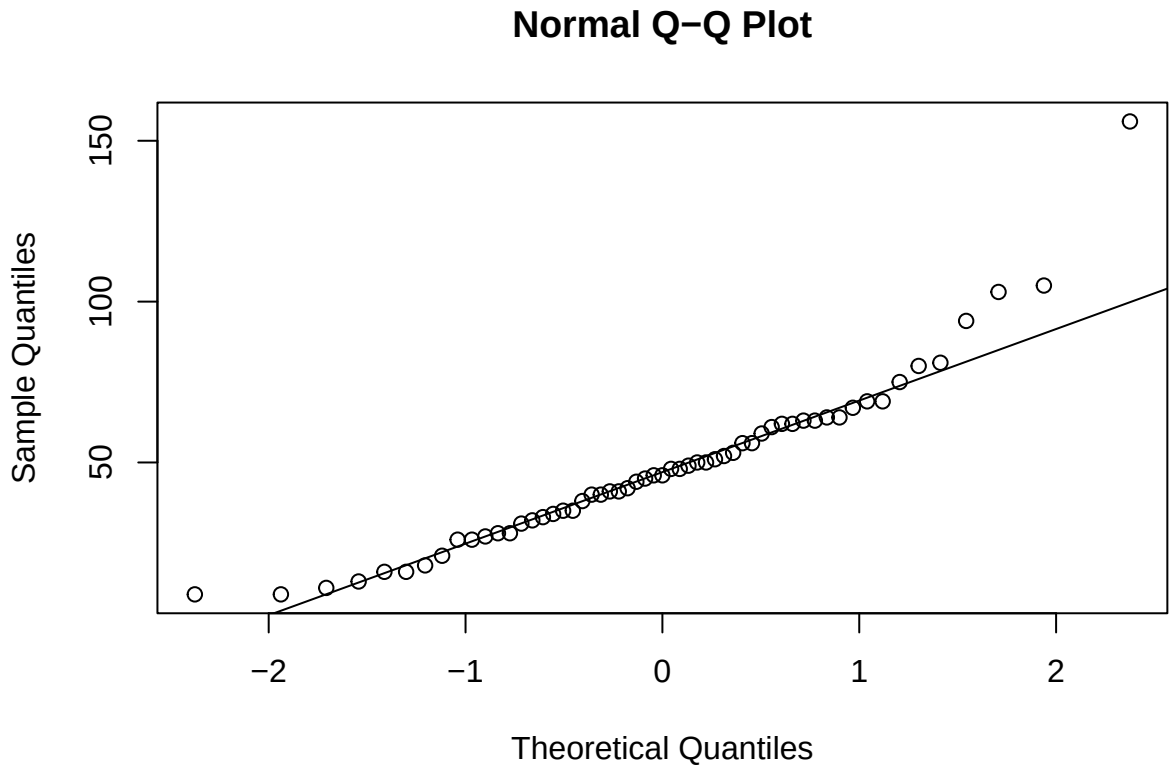
This might be the case because of the different categories of food. Sandwiches probably have a range of sodium, fries probably have a different range of sodium, salads probably have a different range of sodium, and so on. Since different categories of food likely have ranges of sodium that don' overlap much, the pattern appears step-wise.

Exercise 9

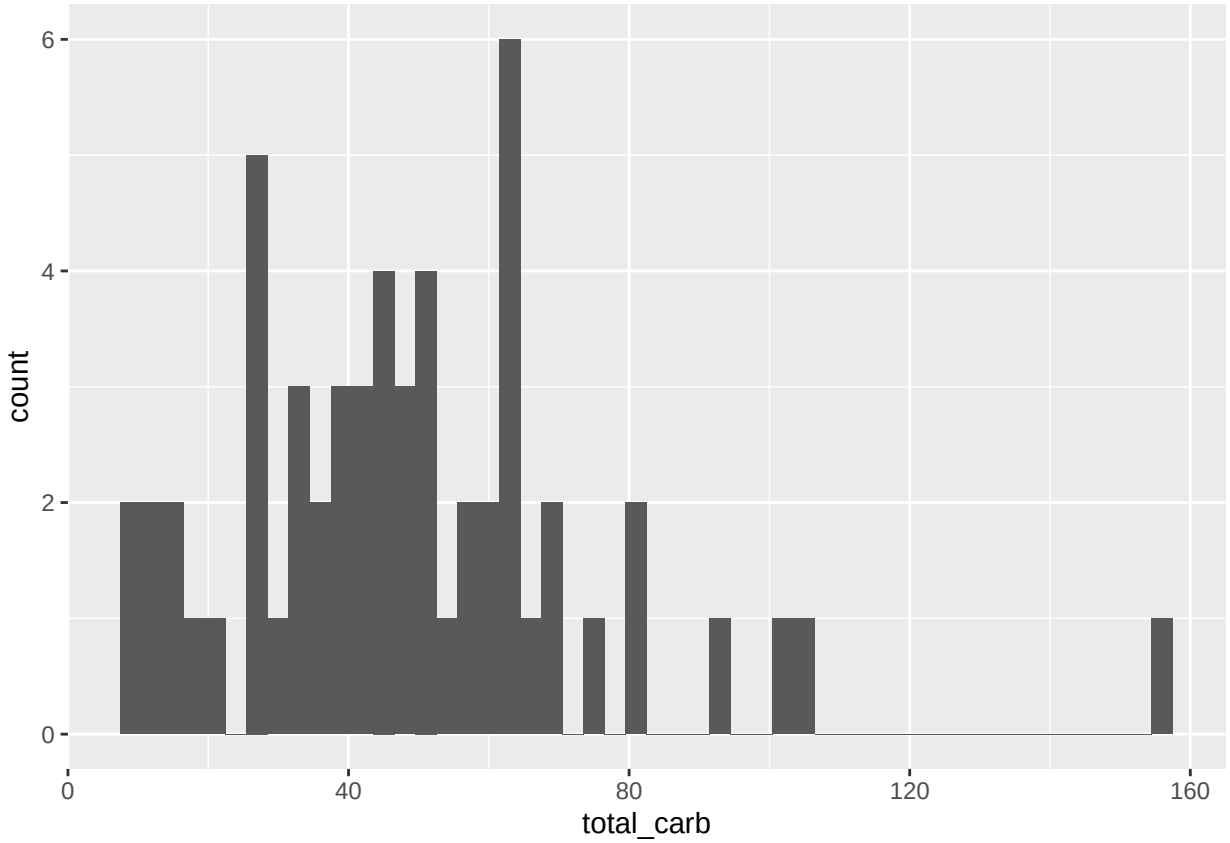
As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

My guess was right-skewed because there seemed to be a tail to the right and above the line in the normal probability plot. The graph confirms this because it is right-skewed with the tail to the right.

```
qqnorm(mcdonalds$total_carb)
qqline(mcdonalds$total_carb)
```




```
ggplot(mcdonalds, aes(x = total_carb)) +
  geom_histogram(bins = 50)
```


6
4
2
0
count
0
40
80
120
160
total_carb