Data 606 Lab 4

Joao De Oliveira

09/30/2025

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

```
data("fastfood", package = "openintro")
head(fastfood)
```

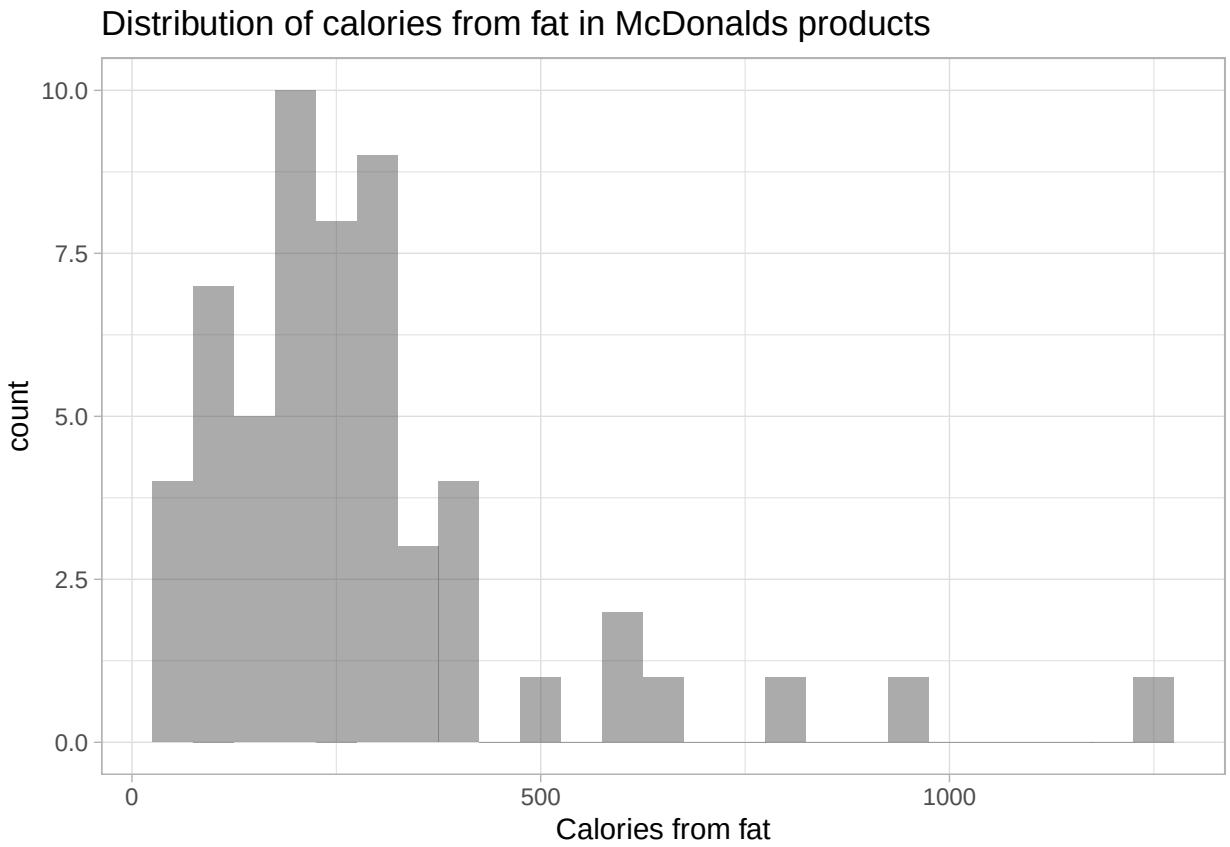
```
## # A tibble: 6 x 17
##   restaurant item       calories cal_fat total_fat sat_fat trans_fat cholesterol
##   <chr>      <chr>         <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
## 1 Mcdonalds  Artisan G~      380      60         7       2       0            95
## 2 Mcdonalds  Single Ba~      840     410        45      17       1.5         130
## 3 Mcdonalds  Double Ba~     1130     600        67      27       3           220
## 4 Mcdonalds  Grilled B~      750     280        31      10       0.5         155
## 5 Mcdonalds  Crispy Ba~      920     410        45      12       0.5         120
## 6 Mcdonalds  Big Mac          540     250        28      10       1            80
## # i 9 more variables: sodium <dbl>, total_carb <dbl>, fiber <dbl>, sugar <dbl>,
## #   protein <dbl>, vit_a <dbl>, vit_c <dbl>, calcium <dbl>, salad <chr>
```

Exercise 1

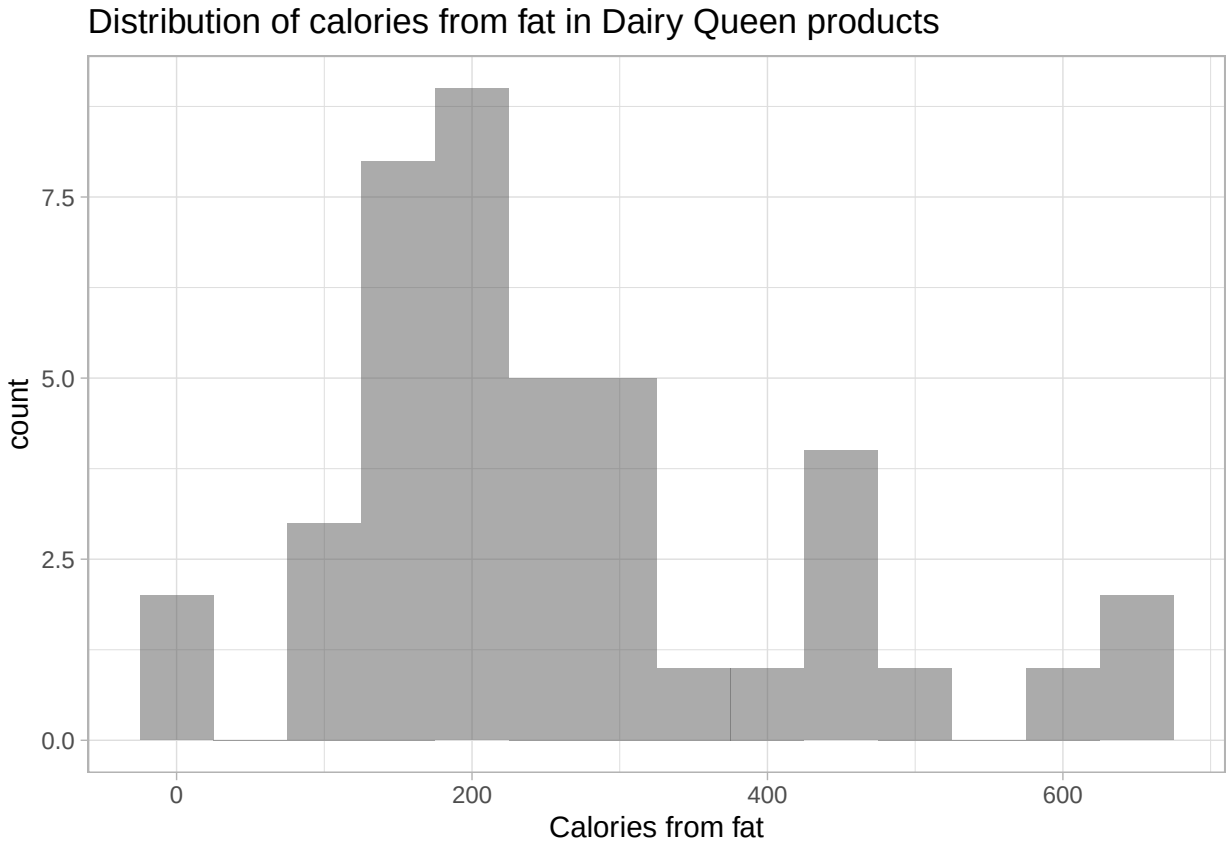
Make a plot (or plots) to visualize the distributions of the amount of calories from fat of the options from these two restaurants. How do their centers, shapes, and spreads compare?

```
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")

ggplot(data = mcdonalds, aes(x = cal_fat)) +
  geom_histogram(binwidth = 50, alpha = 0.5) +
  labs(title = "Distribution of calories from fat in McDonalds products", x = "Calories from fat", y = "
  theme_light()
```




```
ggplot(data = dairy_queen, aes(x = cal_fat)) +
  geom_histogram(binwidth = 50, alpha = 0.5) +
  labs(title = "Distribution of calories from fat in Dairy Queen products", x = "Calories from fat", y =
  theme_light()
```



Answer 1

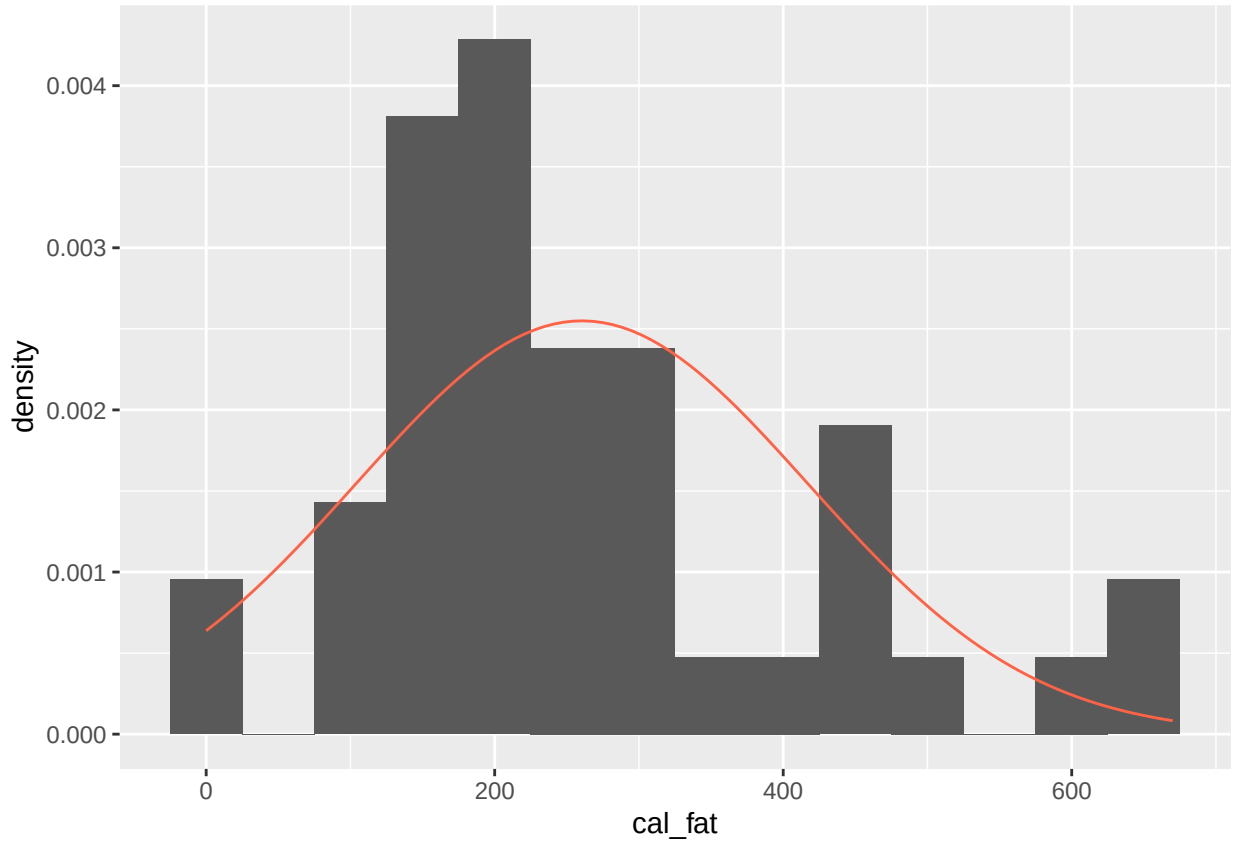
Analyzing McDonalds and Dairy Queen's graphs, we can see that McDonalds' distribution is more spread with some products with more than 750 calories (there is one, which is an outlier, with around 1250 calories) while Dairy Queen's distribution is more narrow, so items are have a more consistent fat distribution. Both distributions are right skewed, since the right tail is much longer, suggesting that both have more low calories products. The center of McDonalds is higher (shifted to the right) when compared with Dairy Queen's center.

Exercise 2

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

```
dqmean <- mean(dairy_queen$cal_fat, na.rm = TRUE)
dqsd <- sd(dairy_queen$cal_fat, na.rm = TRUE)

ggplot(data = dairy_queen, aes(x = cal_fat)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 50) +
  stat_function(fun = dnorm, args = list(mean = dqmean, sd = dqsd), col = "tomato")
```



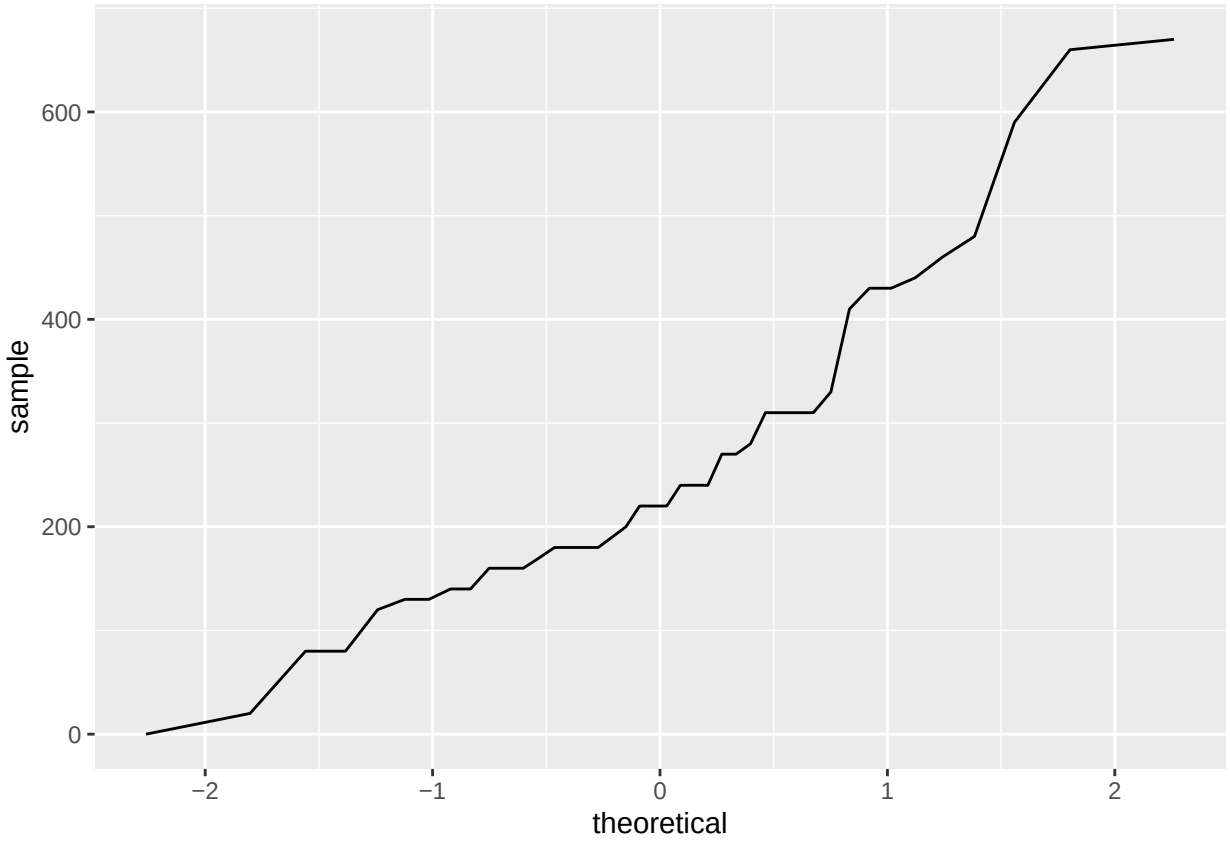
Answer 2

Looking at this plot the distribution doesn't seem to follow a normal distribution since the graph is not really bell shaped despite being closer in some parts, it's pretty irregular due to gaps between columns (low density products with calories from fat between 325 and 425 or none between 525 and 575), and it's not symmetrical. Also, the data is pretty spread to the right side when compared with the left side, reinforcing the idea that this is not a normal distribution.

Exercise 3

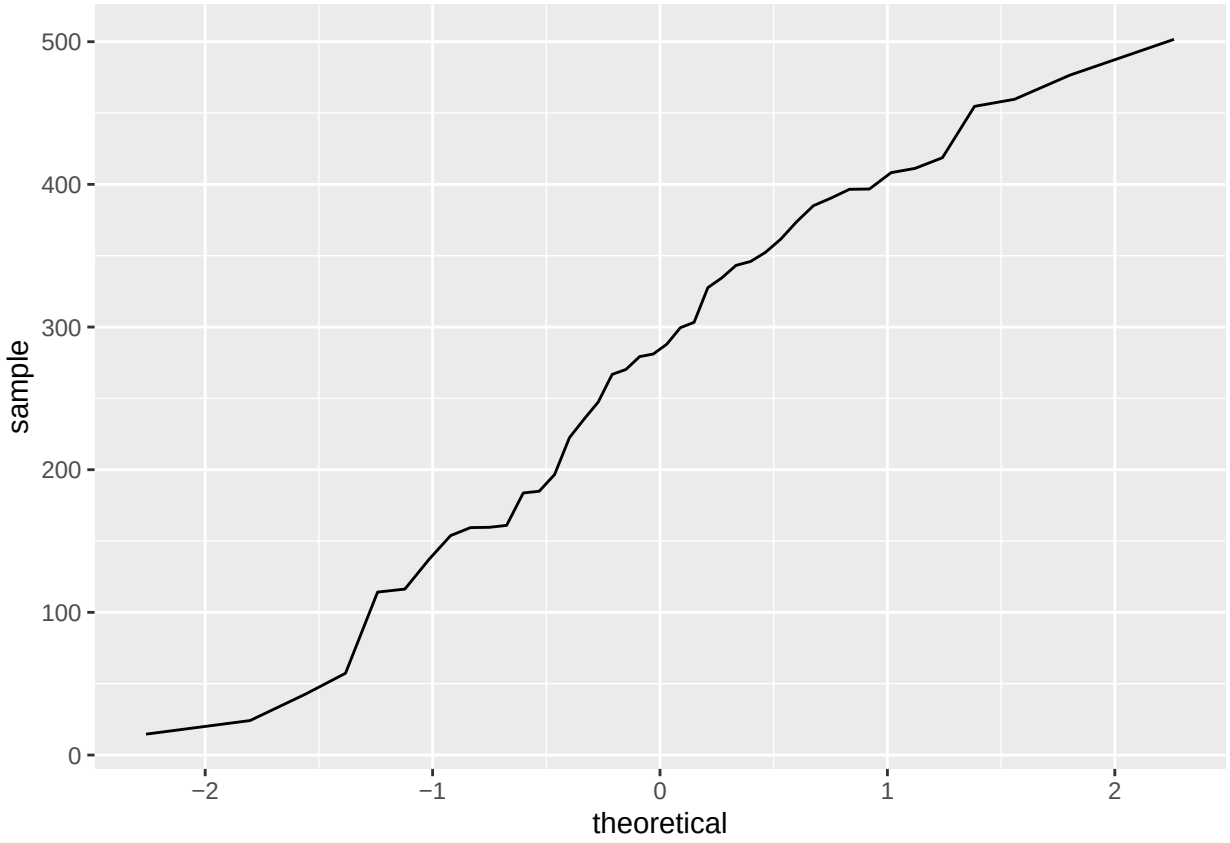
Make a normal probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data? (Since sim_norm is not a data frame, it can be put directly into the sample argument and the data argument can be dropped.)

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




```
sim_norm <- rnorm(n = nrow(dairy_queen), mean = dqmean, sd = dqsd)

ggplot(data = NULL, aes(sample = sim_norm)) +
  geom_line(stat = "qq")
```



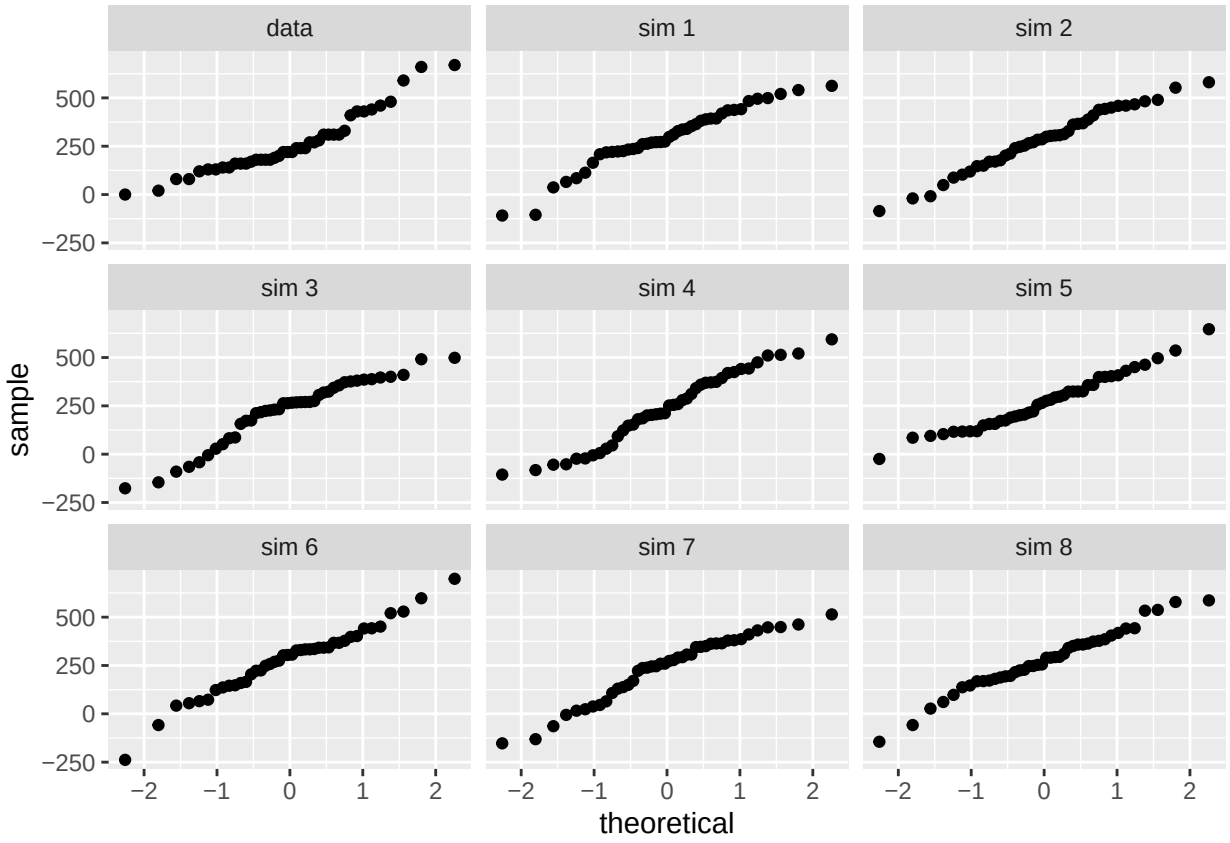
Answer 3

When analyzing the sim_norm probability plot we can see that the points do not all fall on the imaginary digonal (y =x) line. However, we can see that this plot is much closer to a y=x line than the plot with the real data. In the real data plot points tend to have much more deviation from the y=x line. At the same time, in both plots when we get to the top of the plot (products with higher calorie content) the deviation increases.

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?

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



Answer 4

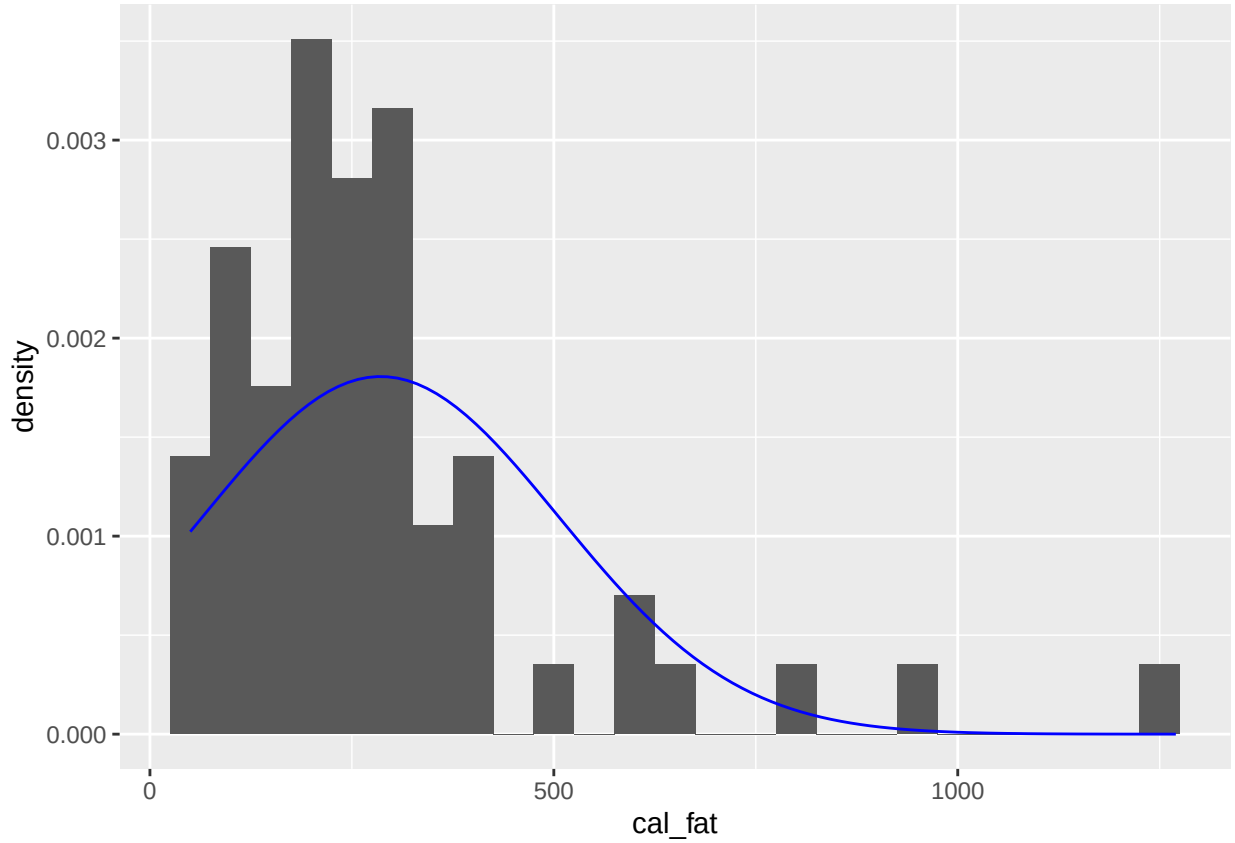
Looking at these 9 plots I conclude that despite they are not equal to the previous simulated plot, they all emphasize the top right corner issue where the distribution becomes more inconsistent. Also, some of them show similar behavior for the bottom left corner portion.

Exercise 5

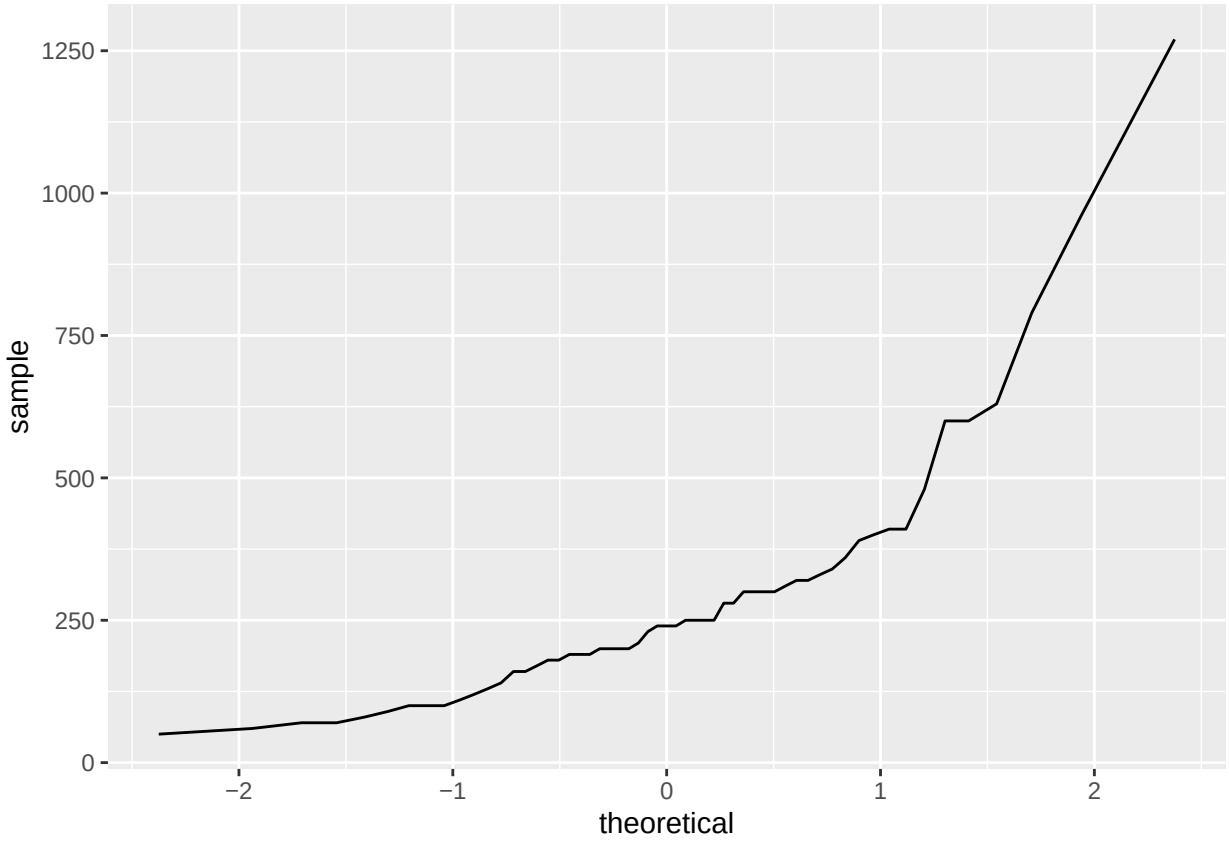
Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

```
mcmean <- mean(mcdonalds$cal_fat, na.rm = TRUE)
mcsd <- sd(mcdonalds$cal_fat, na.rm = TRUE)

ggplot(data = mcdonalds, aes(x = cal_fat)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 50) +
  stat_function(fun = dnorm, args = list(mean = mcmean, sd = mcsd), col = "blue")
```

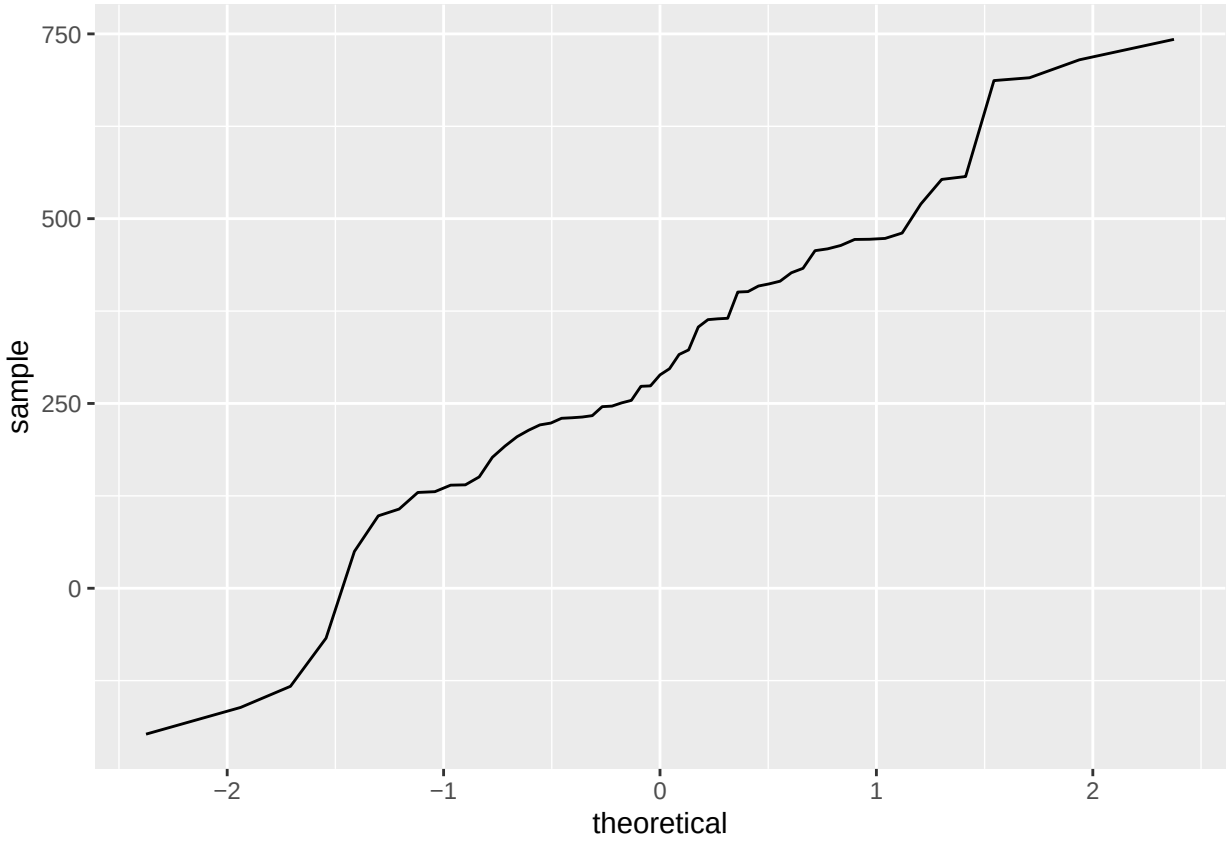



```
ggplot(data = mcdonalds, aes(sample = cal_fat)) +
  geom_line(stat = "qq")
```

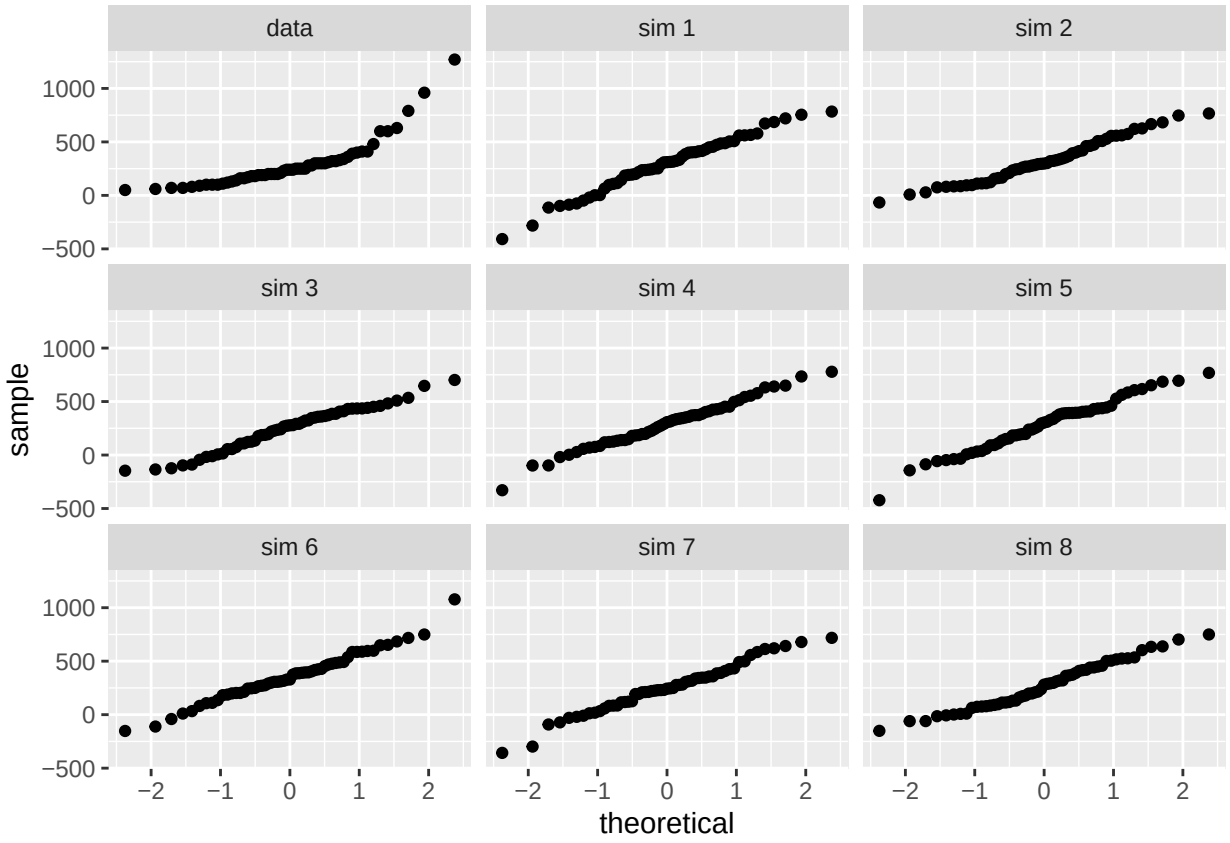



```
sim_norm <- rnorm(n = nrow(mcdonalds), mean = mcmean, sd = mcsd)

ggplot(data = NULL, aes(sample = sim_norm)) +
  geom_line(stat = "qq")
```




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



Answer 5

We can see that Mcdonalds real data distribution is considerable deviated from the diagonal line (y=x). Just like with Dairy Queen products, that deviation is even more obvious for higher calories products (top right corner), which emphasizes the right-skewed distribution. In the simulated data, however, we don't see such an obvious deviation for higher calorie products showing how different it is from the real data. So, Mcdonalds distribution is not normal.

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?

Answer 6

Question 1: What is the probability that an item in the McDonalds menu has less than 200 calories from fat? Question 2: What is the probability of randomly selecting an item from Dairy Queen's menu that has between 250 and 500 calories from fat?

```
# theoretical prob mcdonalds
theoretical_mcdonalds <- pnorm(q = 200, mean = mcmean, sd = mcsd)
print(theoretical_mcdonalds)
```

```
## [1] 0.349167
```

```
# empirical prob mcdonalds
mcdonalds %>%
  filter(cal_fat < 200) %>%
  summarise(percent = n() / nrow(mcdonalds))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.368
```

```
# theoretical prob dairy queen
theoretical_dq <- pnorm(q = 500, mean = dqmean, sd = dqsd) -
  pnorm(q = 250, mean = dqmean, sd = dqsd)
print(theoretical_dq)
```

```
## [1] 0.46376
```

```
# empirical dairy queen
dairy_queen %>%
  filter(cal_fat < 500 & cal_fat > 250) %>%
  summarise(percent = n() / nrow(dairy_queen))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.333
```

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?

```
# check how many restaurants and the sodium summary
unique(fastfood$restaurant)
```

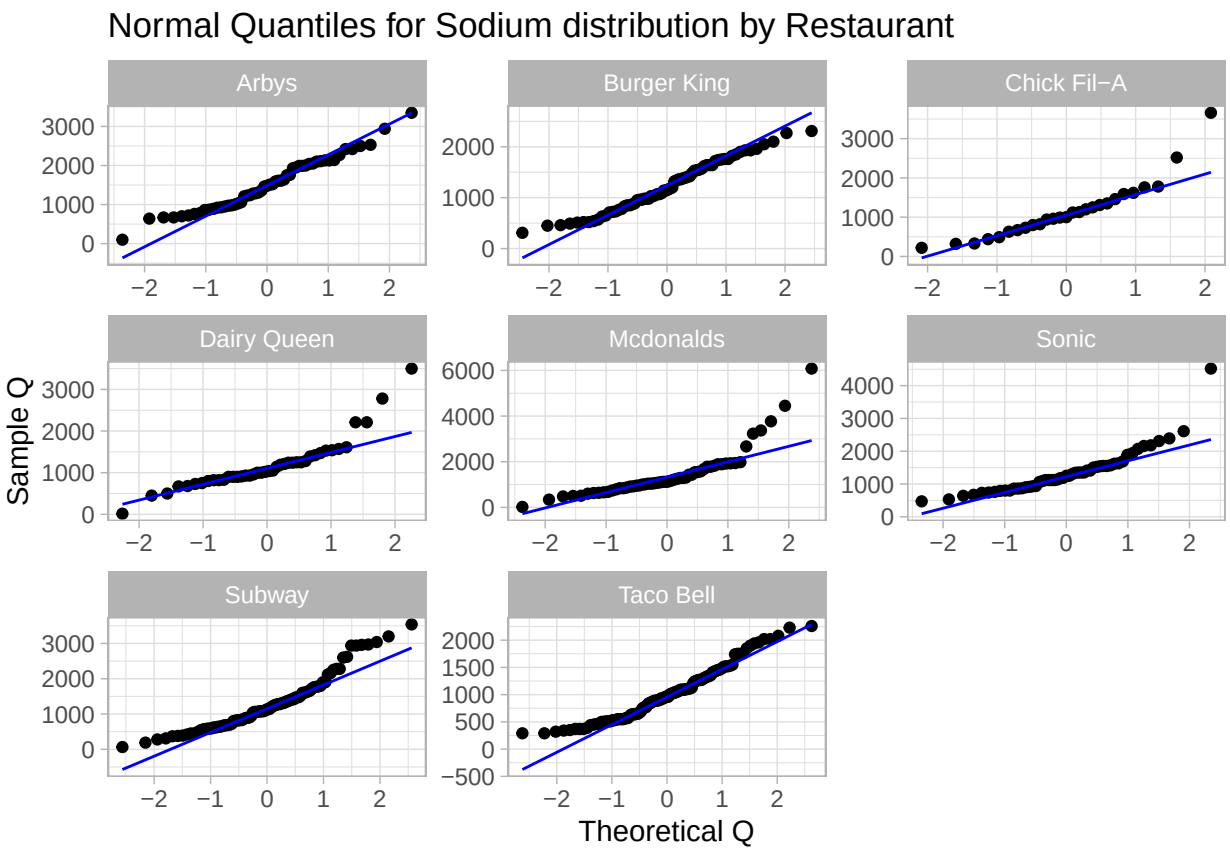
```
## [1] "Mcdonalds"   "Chick Fil-A" "Sonic"       "Arbys"       "Burger King"
## [6] "Dairy Queen" "Subway"      "Taco Bell"
```

```
summary(fastfood$sodium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      15     800    1110    1247    1550    6080
```

```
# plotting the sodium dist for all restaurants
ggplot(fastfood, aes(sample = sodium)) +
  stat_qq() +
  stat_qq_line(color = "blue") +
```

```
facet_wrap(~ restaurant, scales = "free") +
labs(
  title = "Normal Quantiles for Sodium distribution by Restaurant",
  x = "Theoretical Q",
  y = "Sample Q"
) +
theme_light()
```



Answer 7

Looking at the quantile-quantile plots for all restaurants, Arbys and Burger King seem to be the ones' with sodium distribution closer to a normal distribution since the points in their data plots are pretty close to the diagonal line. Chick Fil-A for example has the same behavior for most of the data but then the end tail os quite far apart from the diagonal line, probably showing skeweness.

Exercise 8

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

Answer 8

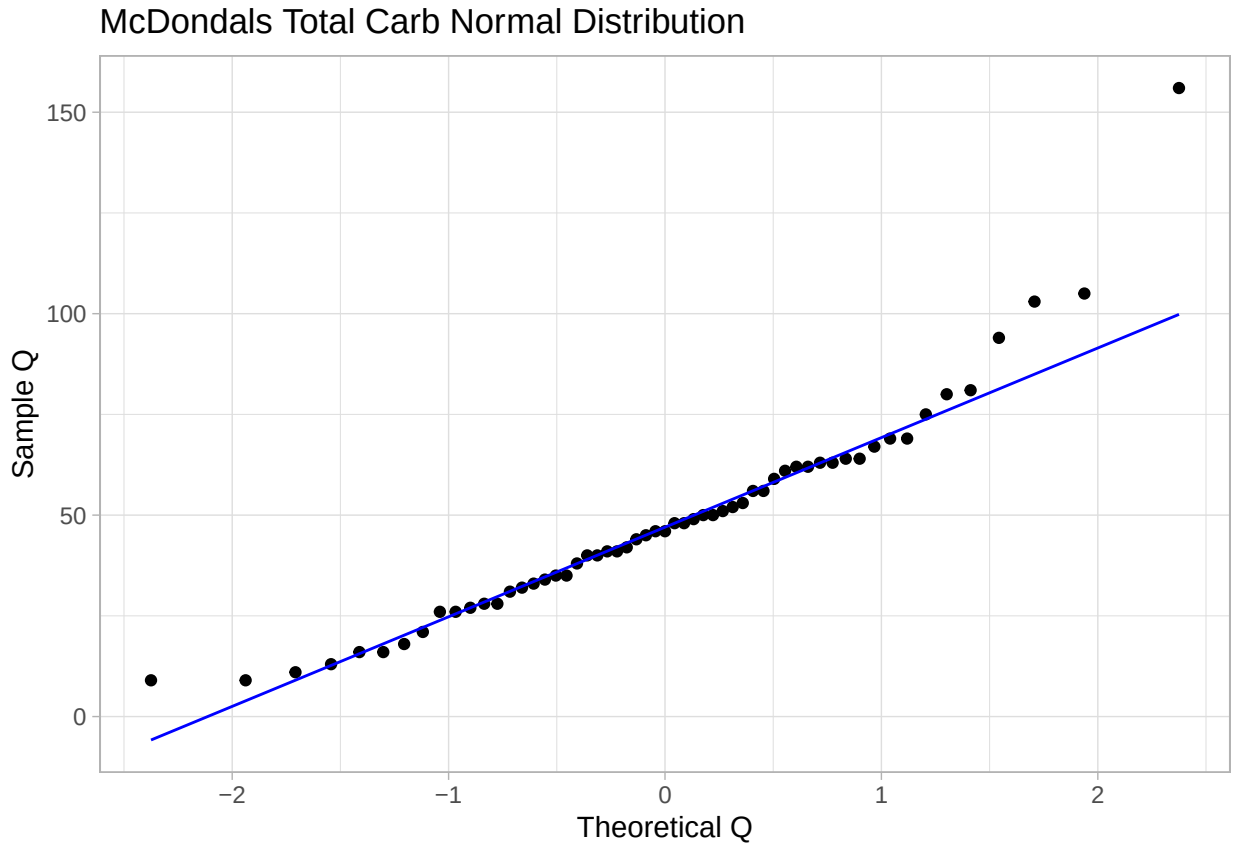
The stepwise pattern seen in the plots has to with the fact that sodium quantities are discrete values, not continuous values. Also, the sample size is not that big, so there's less overlap than if the sample size is

considerably bigger. Finally, in skewed and non normal distributions there will be points considerable above or below the diagonal line, maybe outliers.

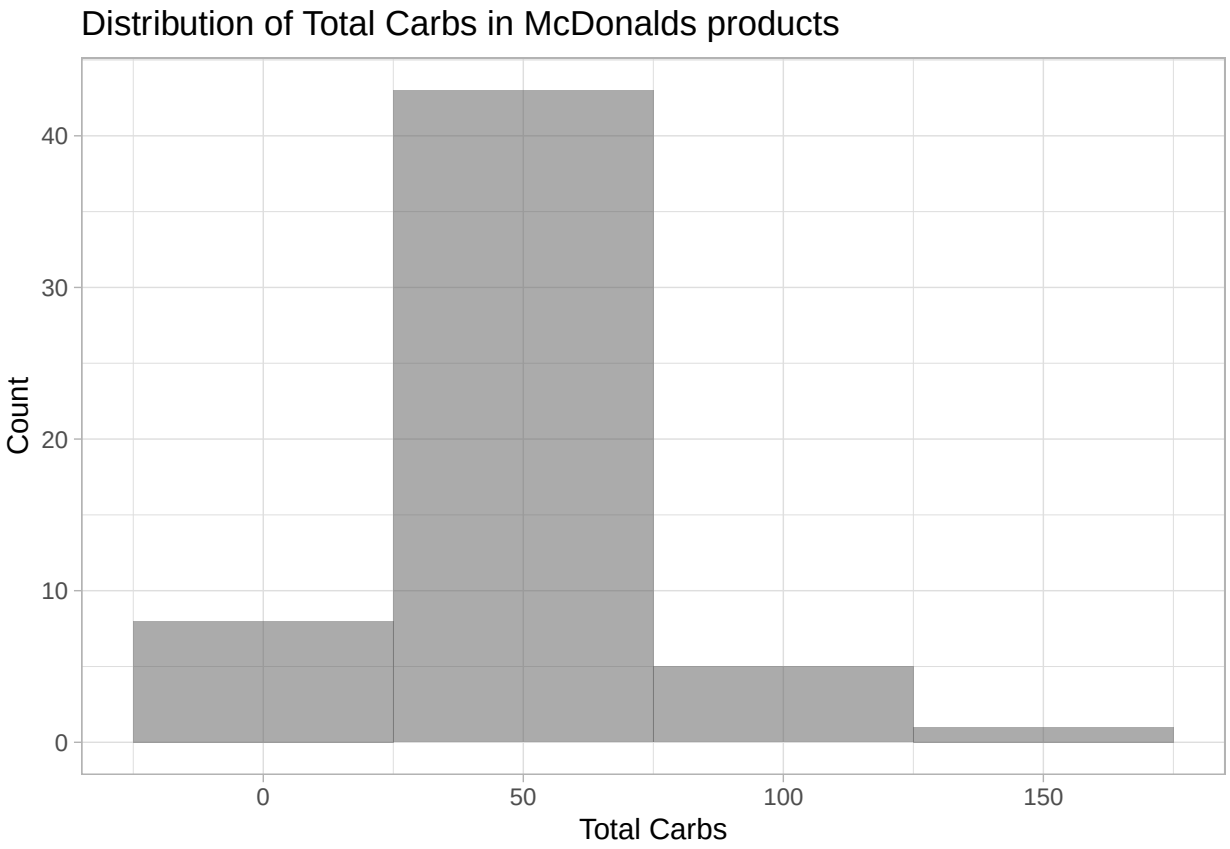
Exercise 9

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.

```
# normal prob plot carbohydrates McDonalds
ggplot(mcdonalds, aes(sample = total_carb)) +
  stat_qq() +
  stat_qq_line(color = "blue") +
  labs(
    title = "McDondals Total Carb Normal Distribution",
    x = "Theoretical Q",
    y = "Sample Q"
  ) +
  theme_light()
```




```
ggplot(data = mcdonalds, aes(x = total_carb)) +
  geom_histogram(binwidth = 50, alpha = 0.5) +
  labs(title = "Distribution of Total Carbs in McDonalds products", x = "Total Carbs", y = "Count") +
  theme_light()
```



Answer 9

The total carbs distribution for McDonalds seems to be right skewed since the normal distribution plot has an upper end tail, the last points are above the diagonal line which suggests right skewness. By looking at the histogram we can confirm that the variable is right skewed since it has a longer tail on the right side (higher carbs).