Lab4 - Normal Distribution

Fomba Kassoh, Colaborated with: Souleymane Doumbia

2023-09-29

Load the required packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

Load the data

Load the fast food data

```
data("fastfood", package='openintro')
fastfood$restaurant <- as.factor(fastfood$restaurant)
head(fastfood, n = 10)
```

```
## # A tibble: 10 x 17
##    restaurant item      calories cal_fat total_fat sat_fat trans_fat cholesterol
##    <fct>      <chr>        <dbl>   <dbl>     <dbl>   <dbl>     <dbl>       <dbl>
##  1 Mcdonalds  Artisan ~      380      60         7       2       0            95
##  2 Mcdonalds  Single B~      840     410        45      17       1.5         130
##  3 Mcdonalds  Double B~     1130     600        67      27       3           220
##  4 Mcdonalds  Grilled ~      750     280        31      10       0.5         155
##  5 Mcdonalds  Crispy B~      920     410        45      12       0.5         120
```

```
##  6 Mcdonalds  Big Mac        540     250        28      10       1            80
##  7 Mcdonalds  Cheesebu~      300     100        12       5       0.5          40
##  8 Mcdonalds  Classic ~      510     210        24       4       0            65
##  9 Mcdonalds  Double C~      430     190        21      11       1            85
## 10 Mcdonalds  Double Q~      770     400        45      21       2.5         175
## # 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>
```

Select Mcdonald and Diary Queen Data

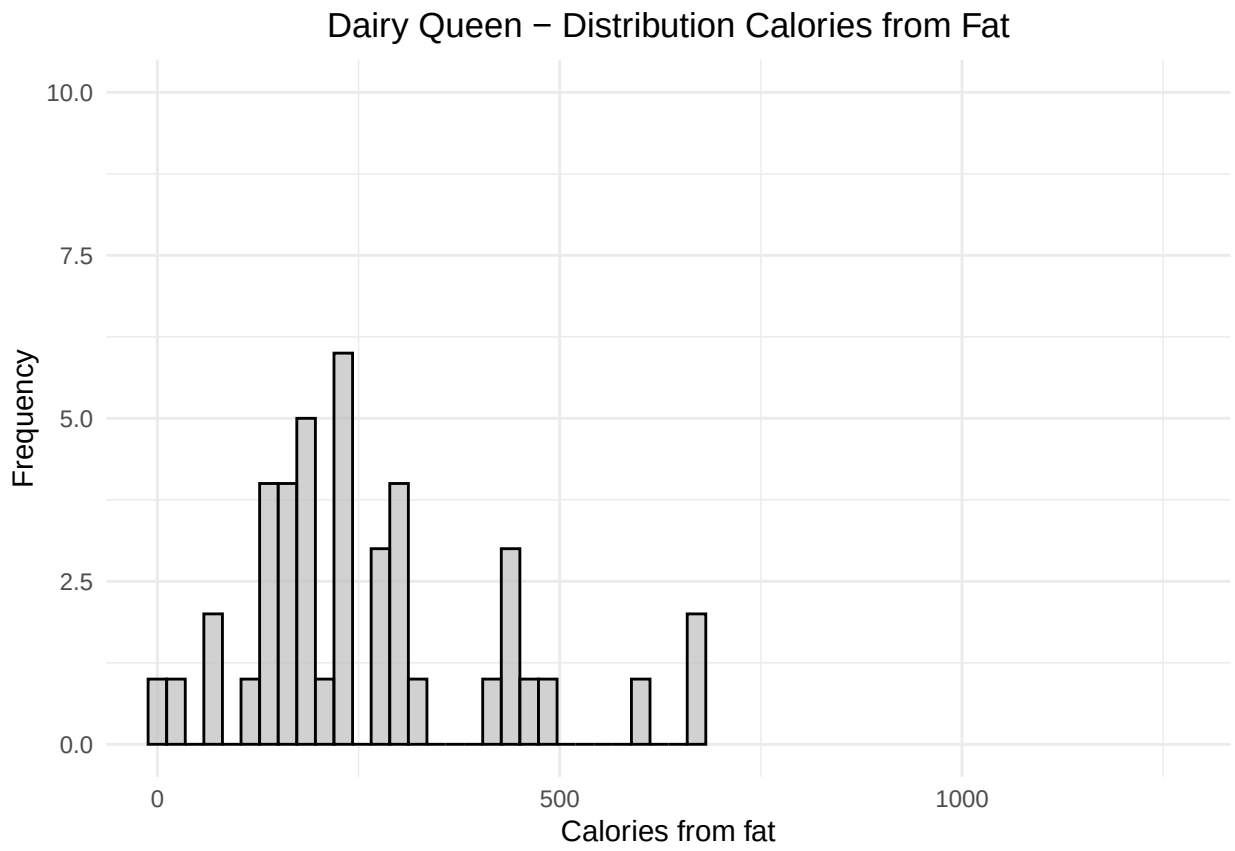
```
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")
```

Excercise 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 descriptions of center, shape, and spread are below the curves.

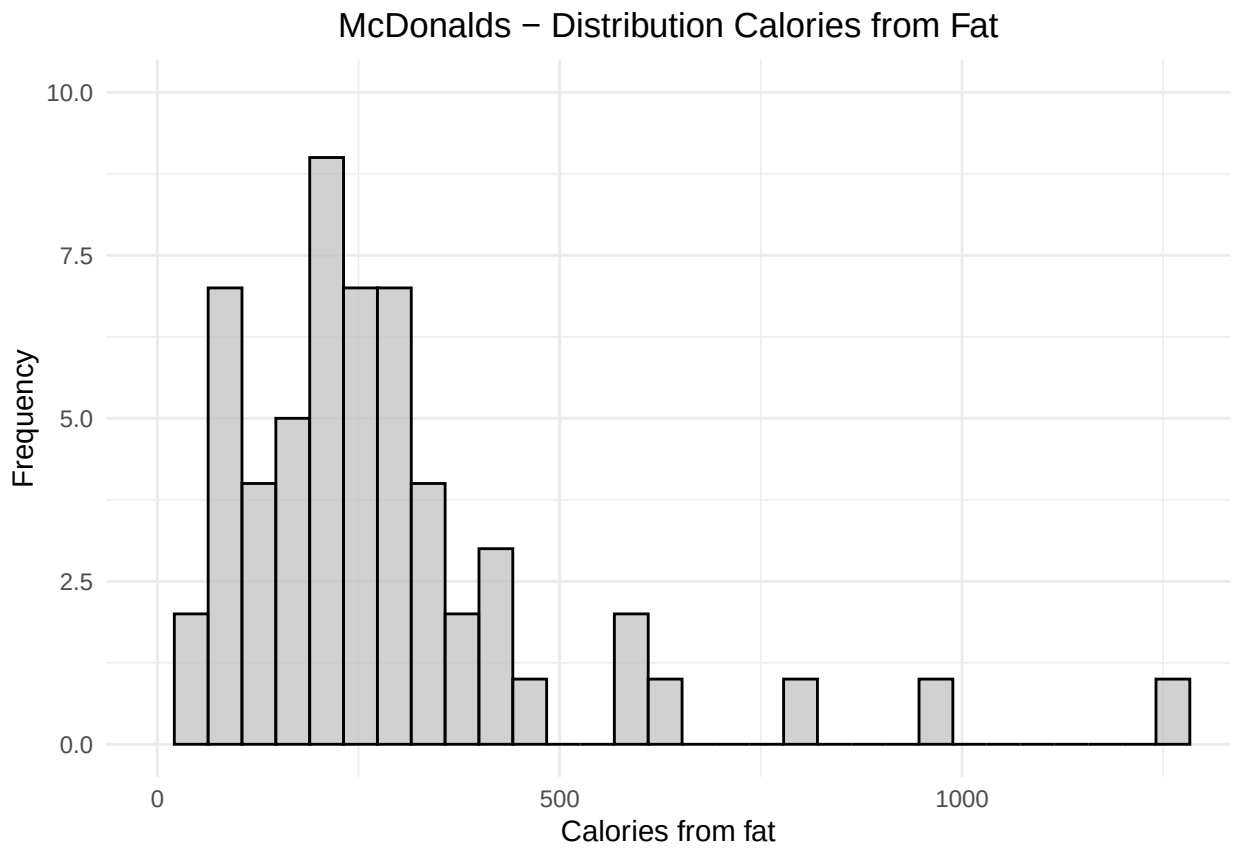
```
# Load the ggplot2 library
library(ggplot2)


diary_queen_plot <- ggplot(dairy_queen, aes(x = cal_fat)) +
  geom_blank() +
  geom_histogram(fill = "grey", color = "black", alpha = 0.7) +
  coord_cartesian(xlim = range(mcdonalds$cal_fat, dairy_queen$cal_fat), ylim = c(0, 10)) +
  labs(title = "Dairy Queen - Distribution Calories from Fat",
       x = "Calories from fat",
       y = "Frequency") +
  theme_minimal()

diary_queen_plot + theme(plot.title = element_text(hjust = 0.5))
```

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




```
mcdonalds_plot <- ggplot(mcdonalds, aes(x = cal_fat)) +
  geom_blank() +
  geom_histogram(fill = "grey", color = "black", alpha = 0.7) +
  coord_cartesian(xlim = range(mcdonalds$cal_fat, dairy_queen$cal_fat), ylim = c(0, 10)) +
  labs(title = "McDonalds - Distribution Calories from Fat",
       x = "Calories from fat",
       y = "Frequency")+
  theme_minimal()

mcdonalds_plot + theme(plot.title = element_text(hjust = 0.5))
```

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

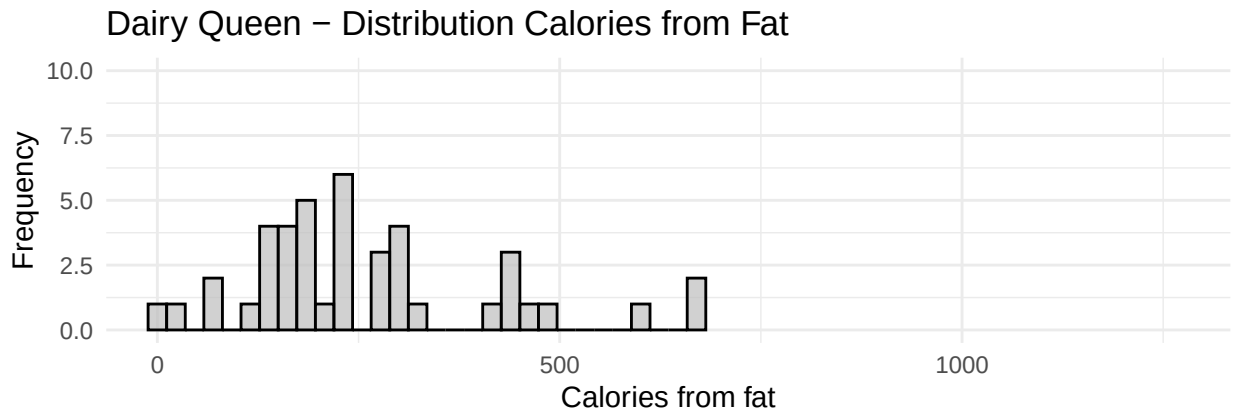


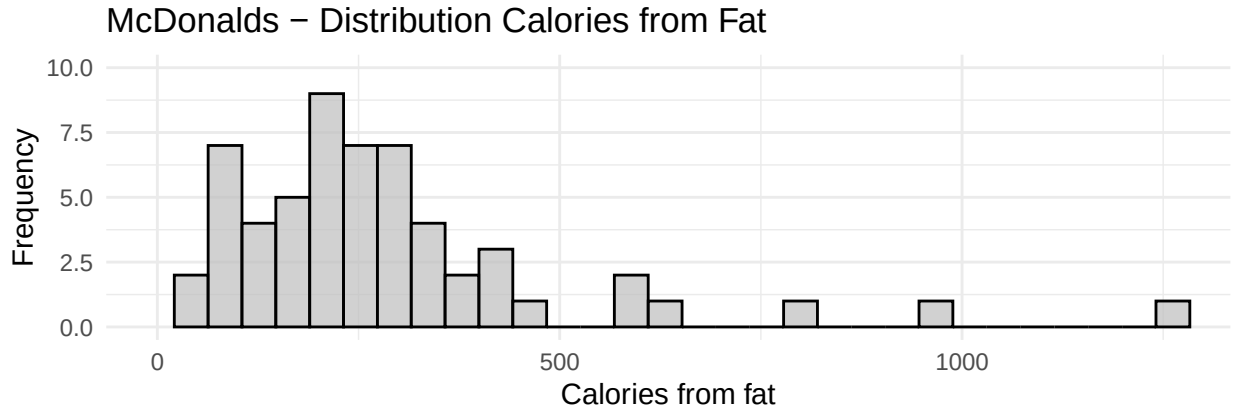

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(diary_queen_plot, mcdonalds_plot, ncol = 1)
```

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





Centers

The peak of the two restuarants are located at about the same calories of fat. McDonalds has the higher peak and therefore has a higher average amount of calories from fat than Diary Queen. The data in McDonalds are more centered around the mean than in Dairy Queen.

Shapes

The two distributions are fairly bell shaped indicating the existence of normality. However, both distributions show a slightly right-skewness

Spreads

Most of the data in McDonalds are around the peak than in Diary Queen. Therefore, calories from fat is more spread out in Diary Queen.

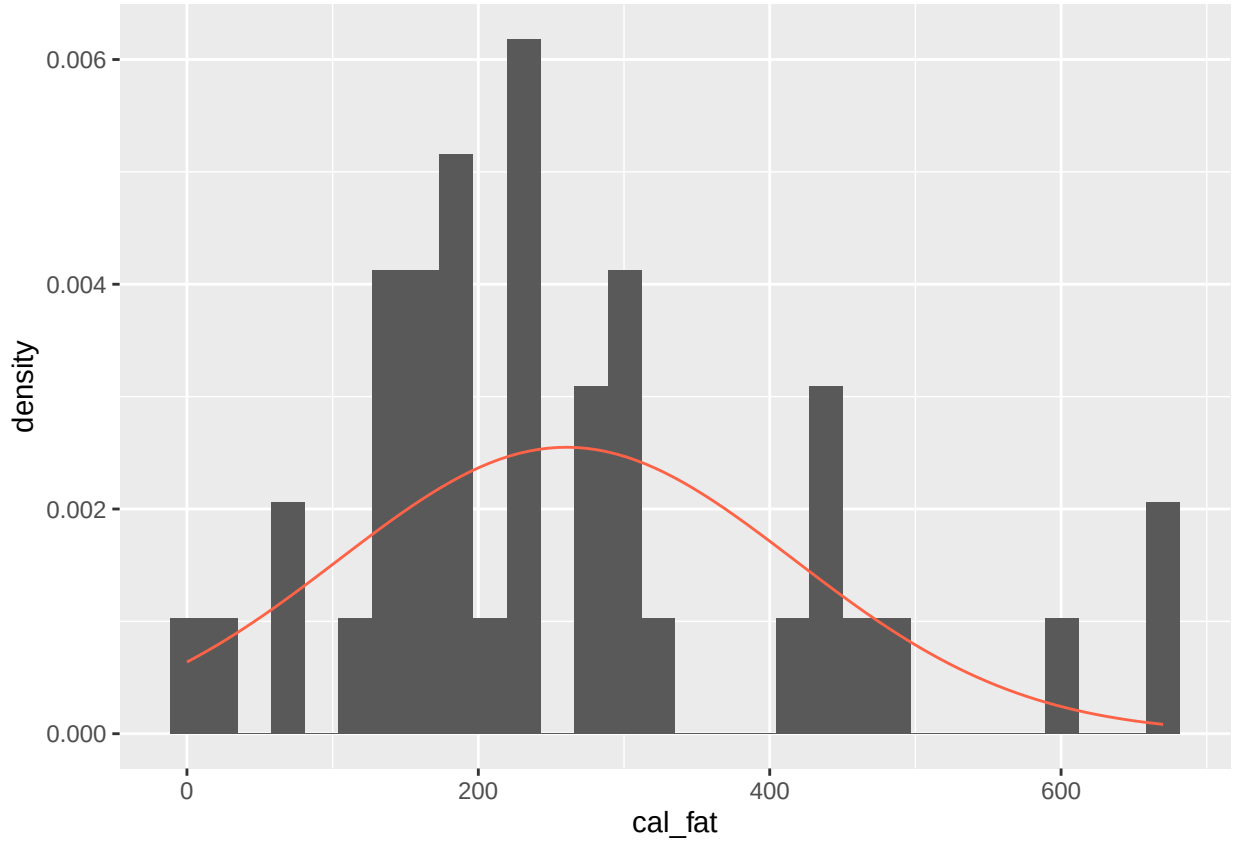
Excercise 2: Based on the this plot, does it appear that the data follow a nearly normal distribution?

Yes. Based on the plot below, the data appears to follow a nearly normal distribution.

```r
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.
```

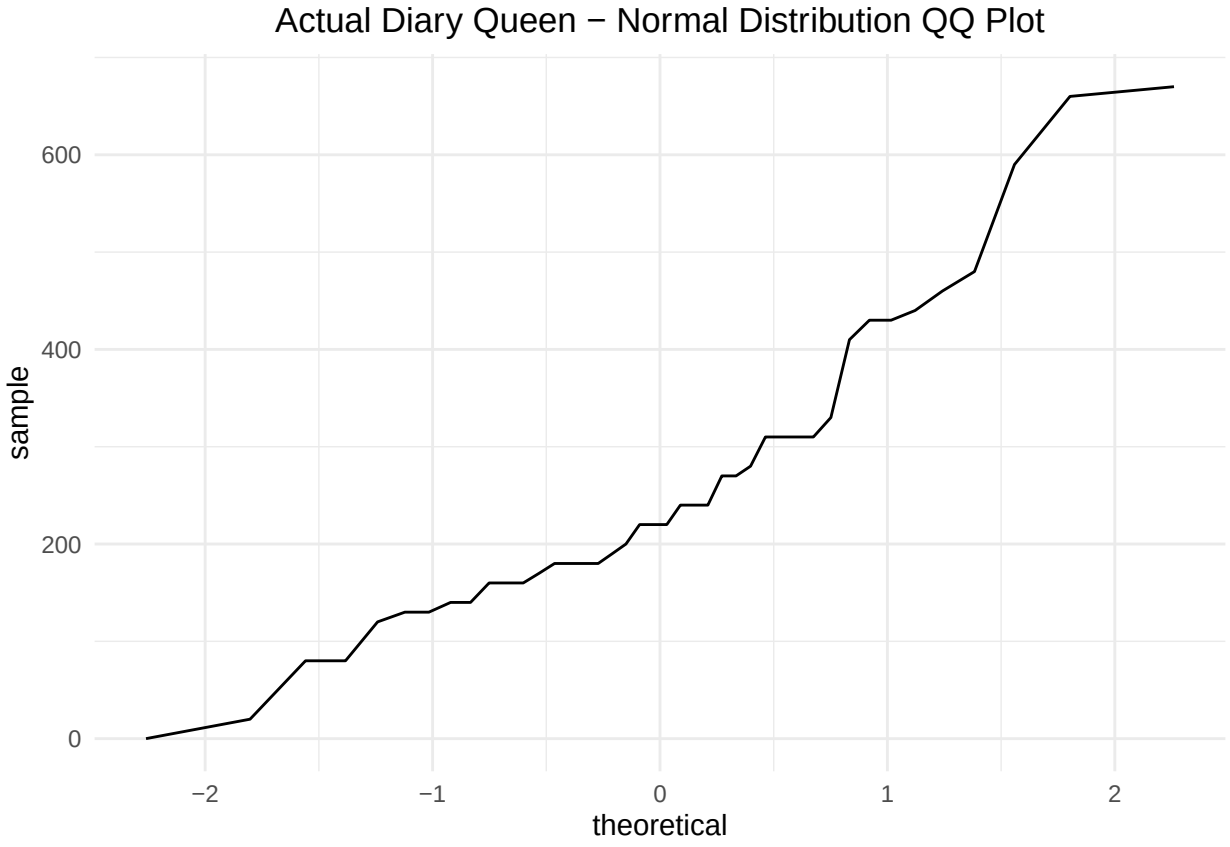
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Excercise 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.)

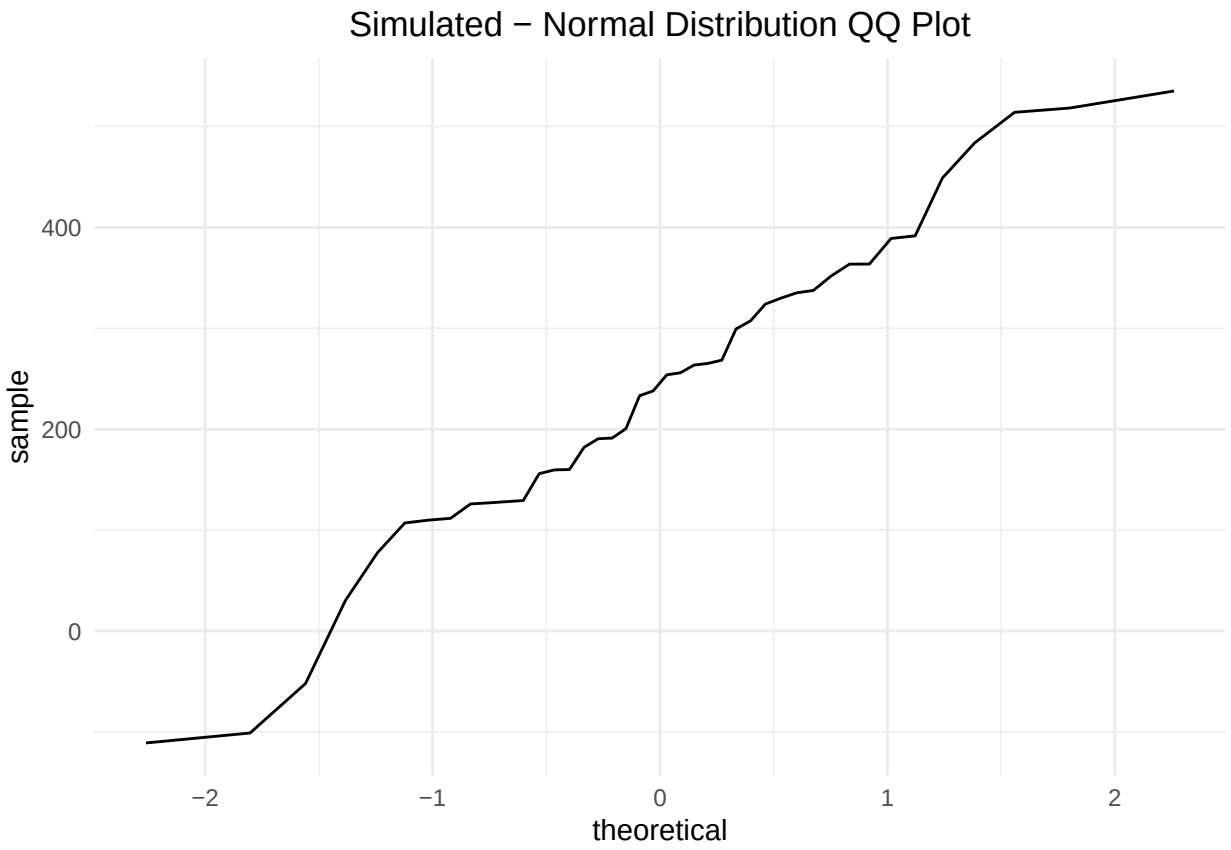
All of the points of the simulated normal distribution sim_norm follows the line. In comparison, the data points of the actual Dairy Queen plot deviate upwards from the 45-degree line which suggests that the data has heavier tails than a normal distribution.

```
ggplot(data = dairy_queen, aes(sample = cal_fat)) +
  geom_line(stat = "qq") +
  labs(title = "Actual Diary Queen - Normal Distribution QQ Plot")+
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



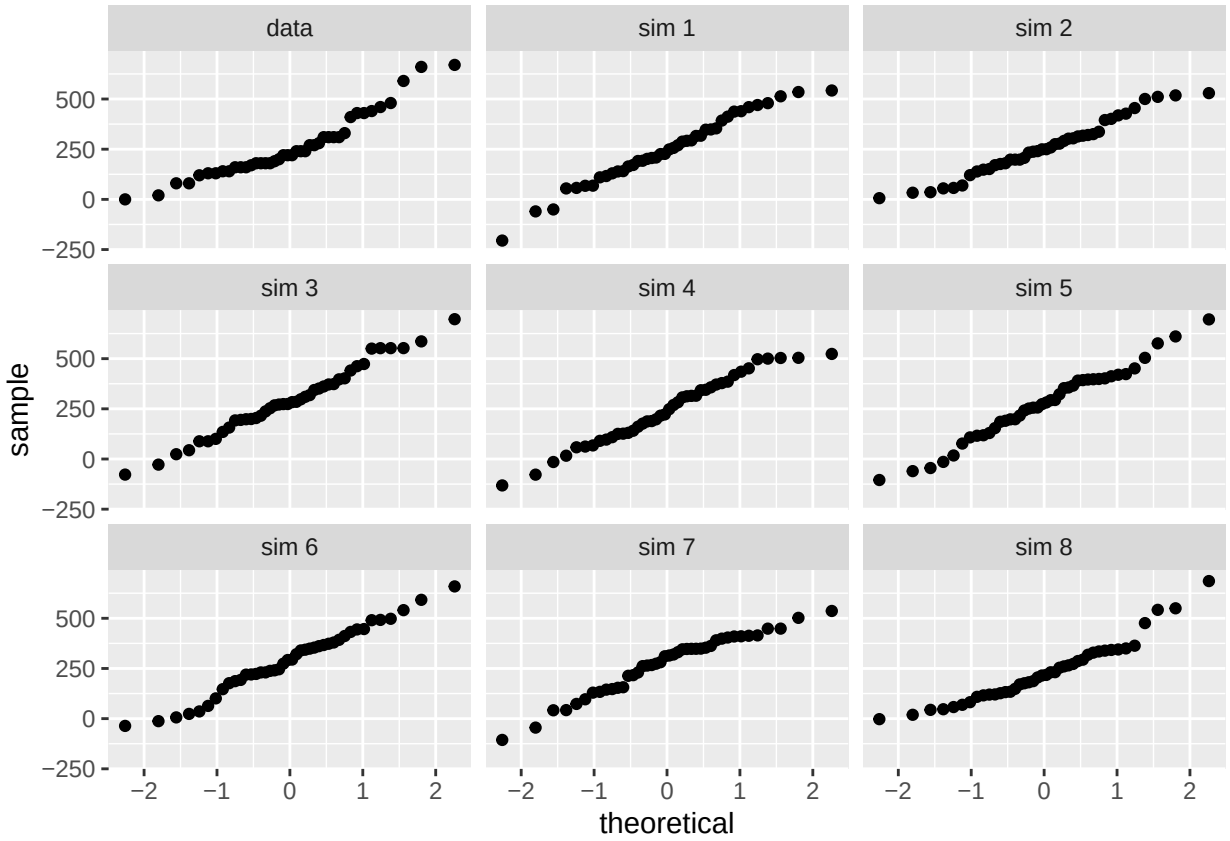

```
sim_norm <- rnorm(n = nrow(dairy_queen), mean = dqmean, sd = dqsd)

#sim_norm_df <- data.frame(x = sim_norm)

ggplot(, aes(sample = sim_norm)) +
  geom_line(stat = "qq") +
  labs(title = "Simulated - Normal Distribution QQ Plot")+
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



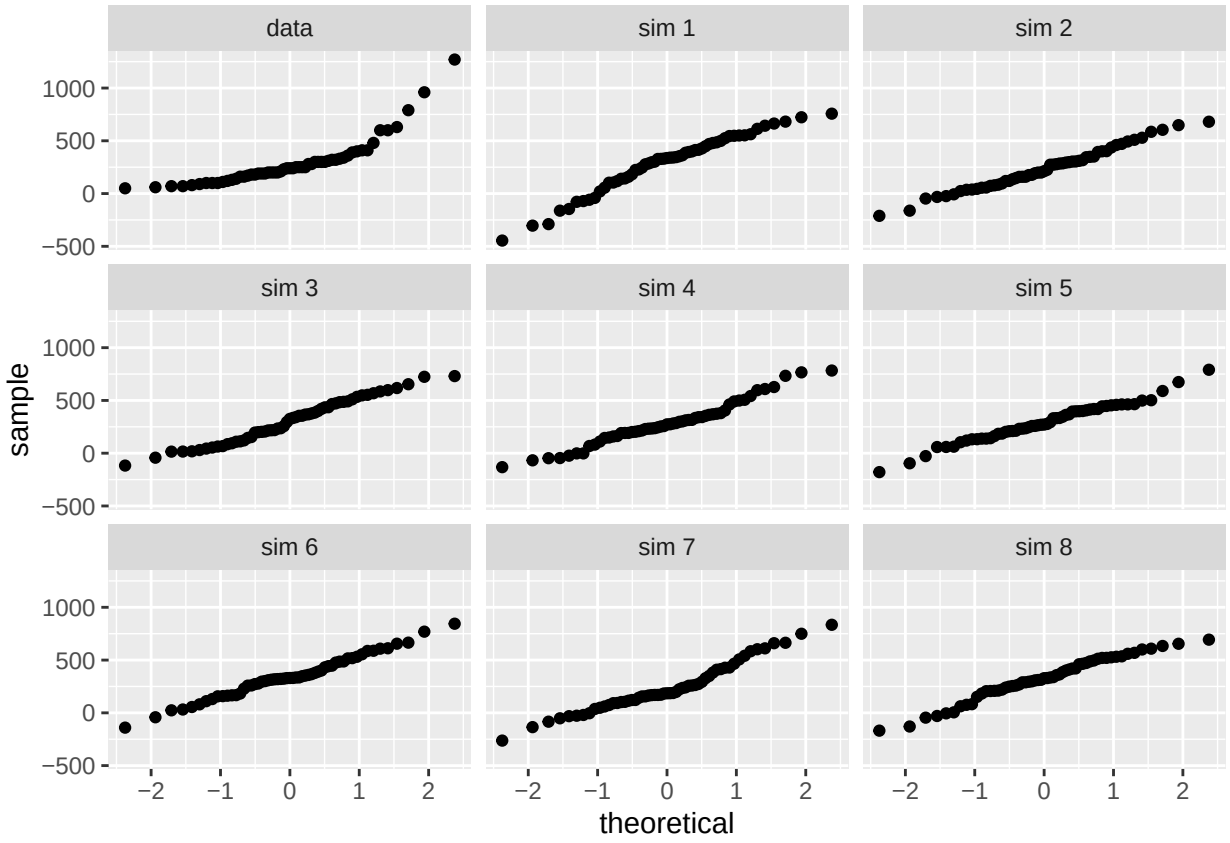
Excercise 4: Does the normal probability plot for the calories from fat look similar to the plots created for the simulated data? That is, do the plots provide evidence that the calories are nearly normal?

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



Excercise 5:Using the same technique, determine whether or not the calories from McDonald's menu appear to come from a normal distribution.

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



Excercise 6:

1. What is the probability that a randomly chosen McDonalds product has less than 300 calories from fat?"
2. What is the probability that a randomly chosen McDonalds product has calories between 100 and 300?"

Based on the results, the theoretical and emperical probabililities for calories between 100 and 300 are closer than the theoretical and emperical probabililities for caloroies from fat less than 300. From the results, the data points for calories are more normally distributed than those for fat calories.

```
mcfat_mean <- mean(mcdonalds$cal_fat)
mcfat_sd   <- sd(mcdonalds$cal_fat)

mcc_mean <- mean(mcdonalds$calories)
mcc_sd   <- sd(mcdonalds$calories)

P_fat_cal_lessthan_300 = pnorm(q = 300, mean = mcfat_mean, sd = mcfat_sd)

p_calories_between_100_and_300 = pnorm(q = 300, mean = mcc_mean, sd = mcc_sd) - (pnorm(q = 100, mean = m

# Calculate emperical probabilities
p_emp_fat_cal_lessthan_300 <- mcdonalds  %>%
                                  filter(cal_fat < 300) %>%
                                  summarise("emp_fat_cal<300" = n() / nrow(mcdonalds))
```

```r
p_emp_p_calories_between_100_and_300 <- mcdonalds  %>%
                                        filter(calories > 100 & calories < 300) %>%
                                        summarise("emp_calories Btw 100 and 300)" = n() / nrow(mcdonal

cbind("fat_cal<300" = P_fat_cal_lessthan_300,
      "emp_calories<300" = p_emp_fat_cal_lessthan_300,
      "calories Btw 100 and 300" = p_calories_between_100_and_300,
      "emp_calories Btw 100 and 300" = p_emp_p_calories_between_100_and_300
)
```

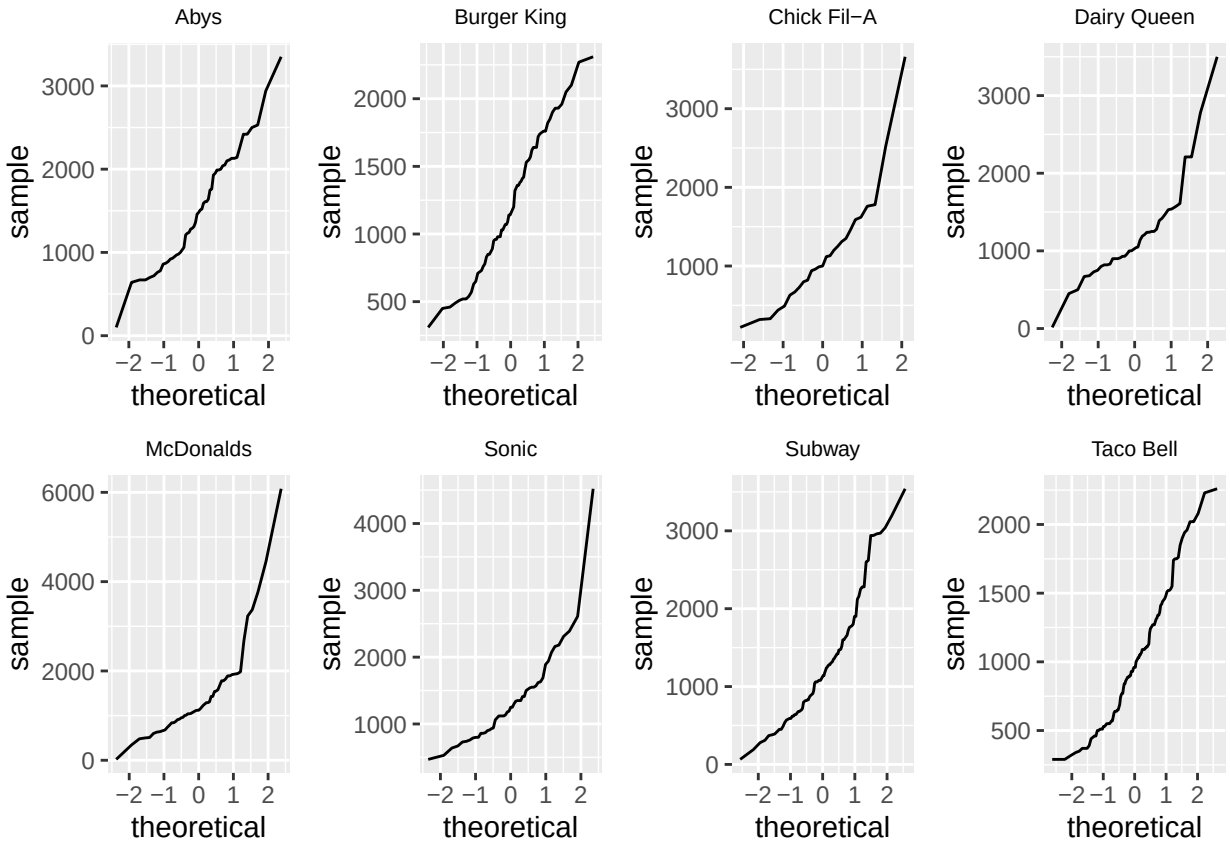
```
##   fat_cal<300 emp_fat_cal<300 calories Btw 100 and 300
## 1   0.5259626       0.6315789                0.1094941
##   emp_calories Btw 100 and 300)
## 1                     0.1403509
```

Excercise 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?

Of the different restaurants, the distribution is closest to normal for sodium for Burger King, Abys, Taco Bell, and Subway in that order.

```r
arbys <- fastfood %>%
  filter(restaurant == "Arbys")

burger_king <- fastfood %>%
  filter(restaurant == "Burger King")

chick_fil_A <- fastfood %>%
  filter(restaurant == "Chick Fil-A")

sonic <- fastfood %>%
  filter(restaurant == "Sonic")

subway <- fastfood %>%
  filter(restaurant == "Subway")

taco_bell <- fastfood %>%
  filter(restaurant == "Taco Bell")



arbys_plot <- ggplot(data = arbys, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Abys")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

burger_King_plot <- ggplot(data = burger_king, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Burger King")+
```

```
  theme(plot.title = element_text(size = 8, hjust = 0.5))

chick_Fil_A_plot <- ggplot(data = chick_fil_A, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Chick Fil-A")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

dary_queen_plot <- ggplot(data = dairy_queen, aes(sample = sodium)) +
  geom_line(stat = "qq")+
  labs(title = "Dairy Queen")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

mcdonals_plot <- ggplot(data = mcdonalds, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "McDonalds")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

sonic_plot <- ggplot(data = sonic, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Sonic")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

subway_plot <- ggplot(data = subway, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Subway")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

taco_bell <- ggplot(data = taco_bell, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  labs(title = "Taco Bell")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

library(gridExtra)

grid.arrange(arbys_plot,
             burger_King_plot,
             chick_Fil_A_plot,
             dary_queen_plot,
             mcdonals_plot,
             sonic_plot,
             subway_plot,
             taco_bell,
             ncol = 4)
```



Excercise 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?

The stepwise pattern in the Q-Q plot for sodium distributions suggests that the data may have some discrete characteristics rather than being continuously distributed.It may also be due to outliers in the data.

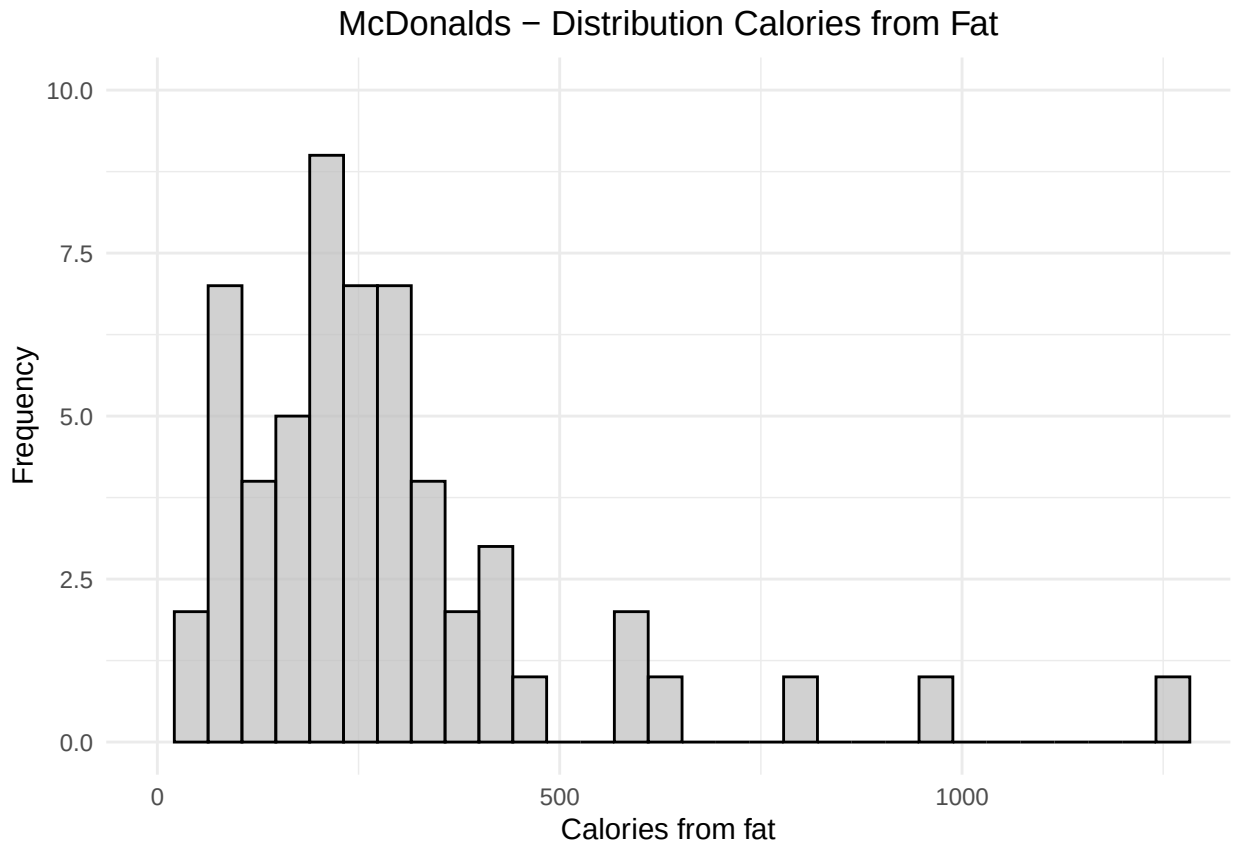
Excercise 9: As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

Q-Q plot points diverge from a straight line which suggests skewness. The QQ-Plot bends downward which indicates right-skewness. This is confirmed by the histogram plot.

```
mcdonals_qq_plot <- ggplot(data = mcdonalds, aes(sample = total_carb)) +
  geom_line(stat = "qq") +
  labs(title = "McDonalds QQ Plot")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))

mcdonalds_hist_plot <- ggplot(mcdonalds, aes(x = cal_fat)) +
  geom_blank() +
```

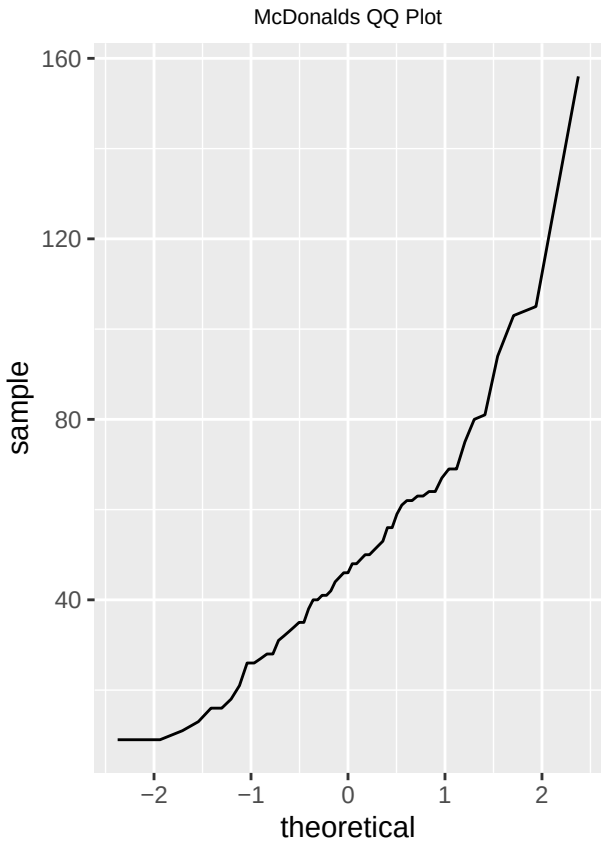
```
  geom_histogram(fill = "grey", color = "black", alpha = 0.7) +
  labs(title = "McDonalds Histogram")+
  theme(plot.title = element_text(size = 8, hjust = 0.5))


mcdonalds_plot + theme(plot.title = element_text(hjust = 0.5))
```

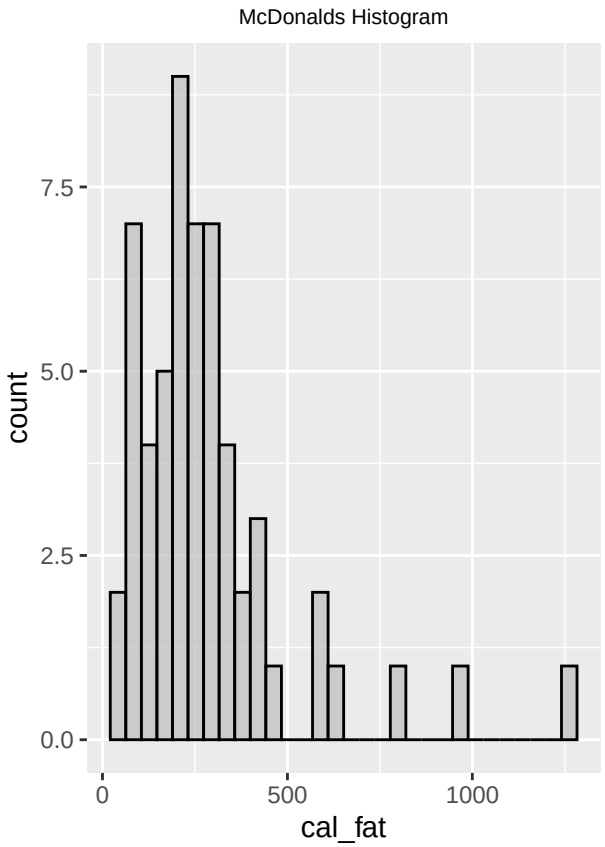
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
library(gridExtra)

grid.arrange(mcdonals_qq_plot, mcdonalds_hist_plot, ncol = 2)
```

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


McDonalds QQ Plot
160
120
80
40
sample
−2
−1
0
1
2
theoretical


McDonalds Histogram
7.5
5.0
2.5
0.0
count
0
500
1000
cal_fat