Lab 4

Alice Ding

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

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

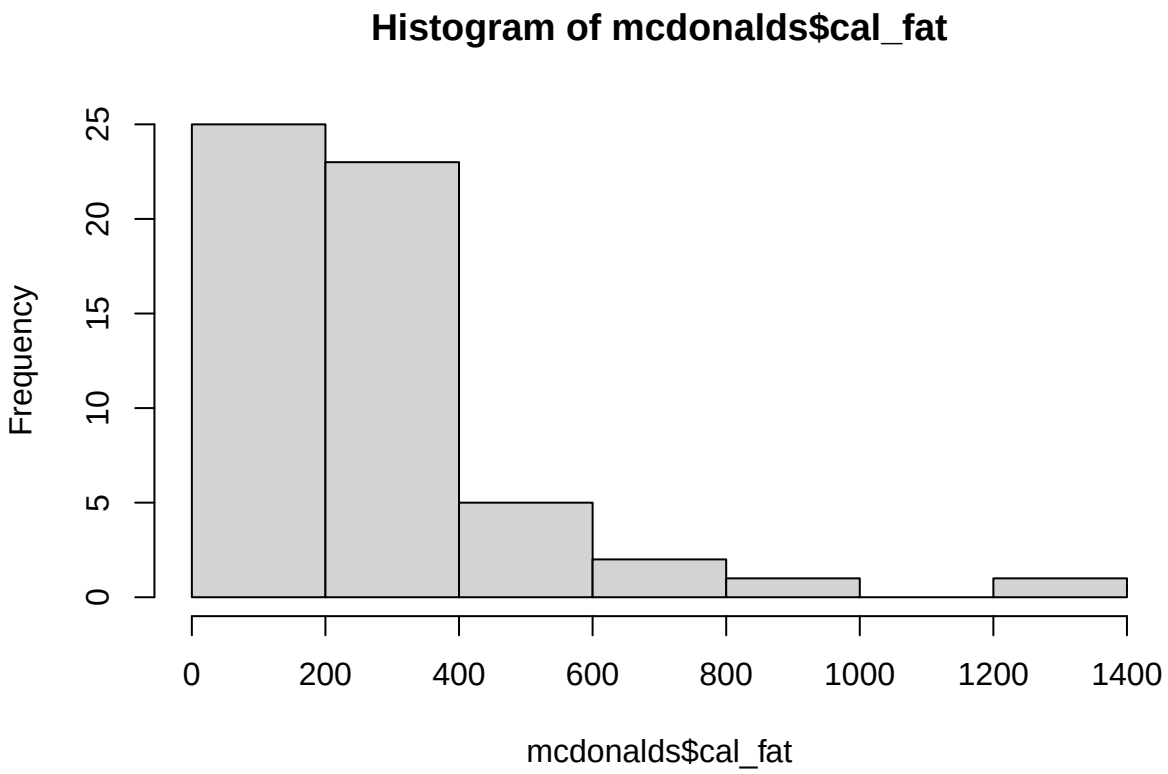
```
library(ggplot2)
library(dplyr)

data("fastfood", package='openintro')
mcdonalds <- fastfood %>%
  filter(restaurant == "Mcdonalds")
dairy_queen <- fastfood %>%
  filter(restaurant == "Dairy Queen")
```

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?

```
hist(mcdonalds$cal_fat)
```




```
hist(dairy_queen$cal_fat)
```

McDonald's is very much right skewed with a majority of their calories from fat ranging from 0 to 400 calories. The center of the data would likely lie around 200 while the data goes from 0 to 1400 calories, however things towards the left side of that data are very, very sparse.

Dairy Queen has a much more balanced range as it's not as extreme as McDonald's with how far right it skews. A majority of foods seem to be in the 100-200 calorie range, however the taper is not as apparent. The center of this data will likely be a little higher, perhaps between 200 and 300, while the data only goes from 0 to 700 calories.

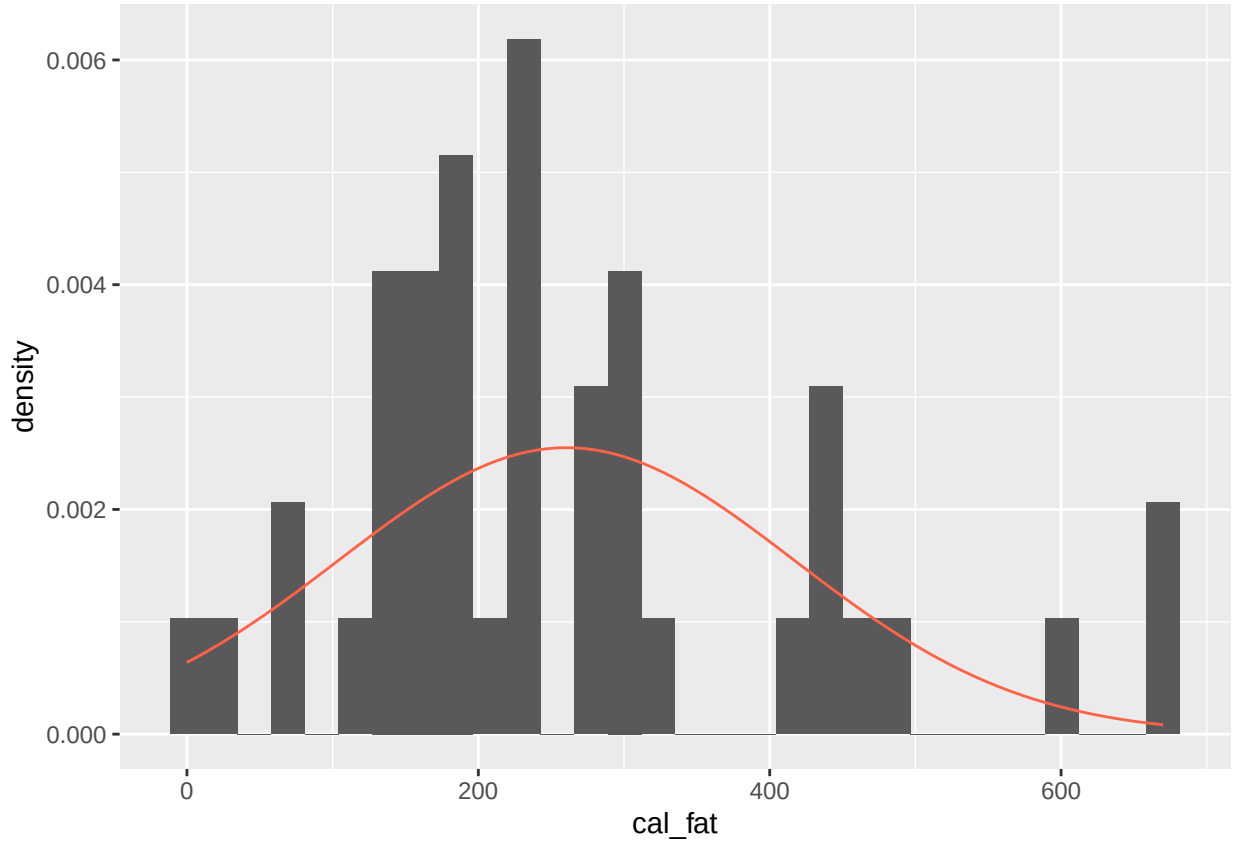
Exercise 2

Based on the this plot, does it appear that the data follow a nearly 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.
```

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

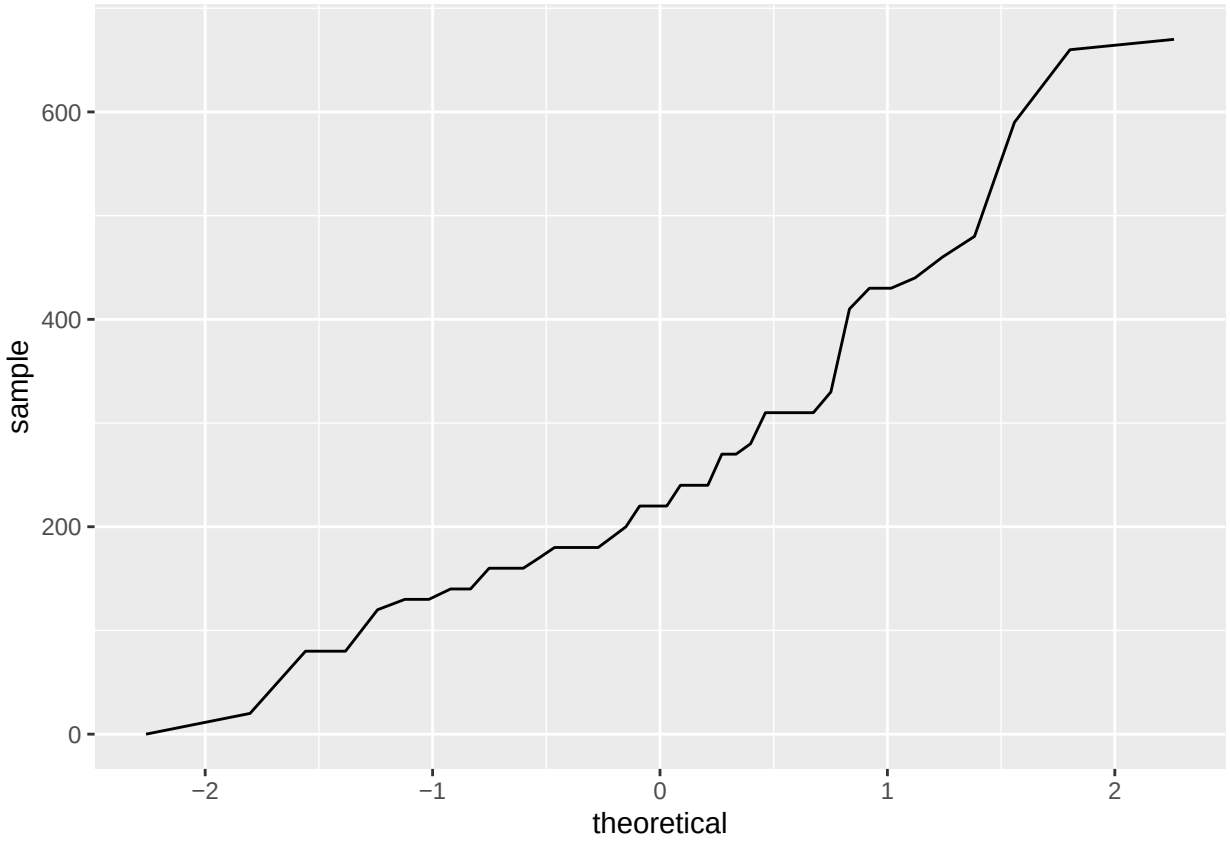


It does look relatively normal, however the peaks of the data are more extreme than the line depicts. All in all though, the data does follow the trend of the bell curve very well.

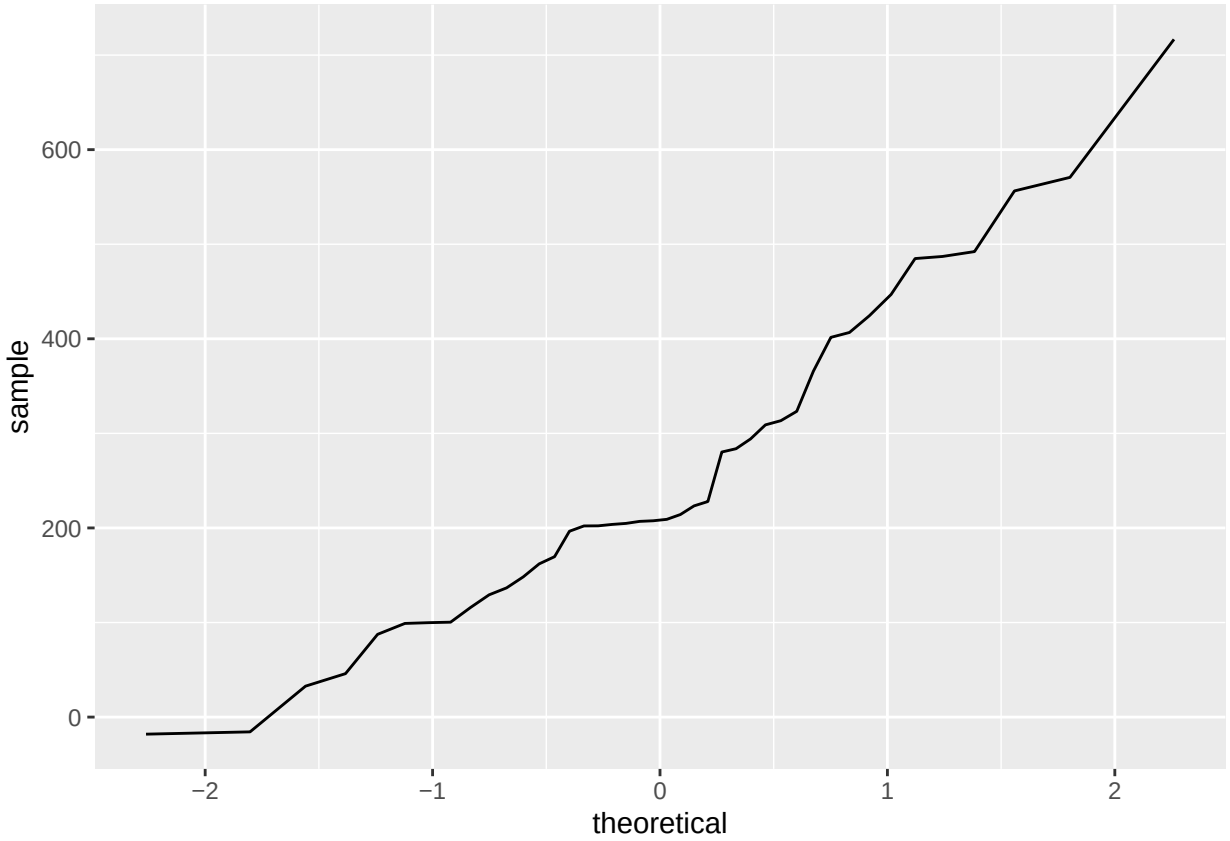
Exercise 3

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?

```
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")
```



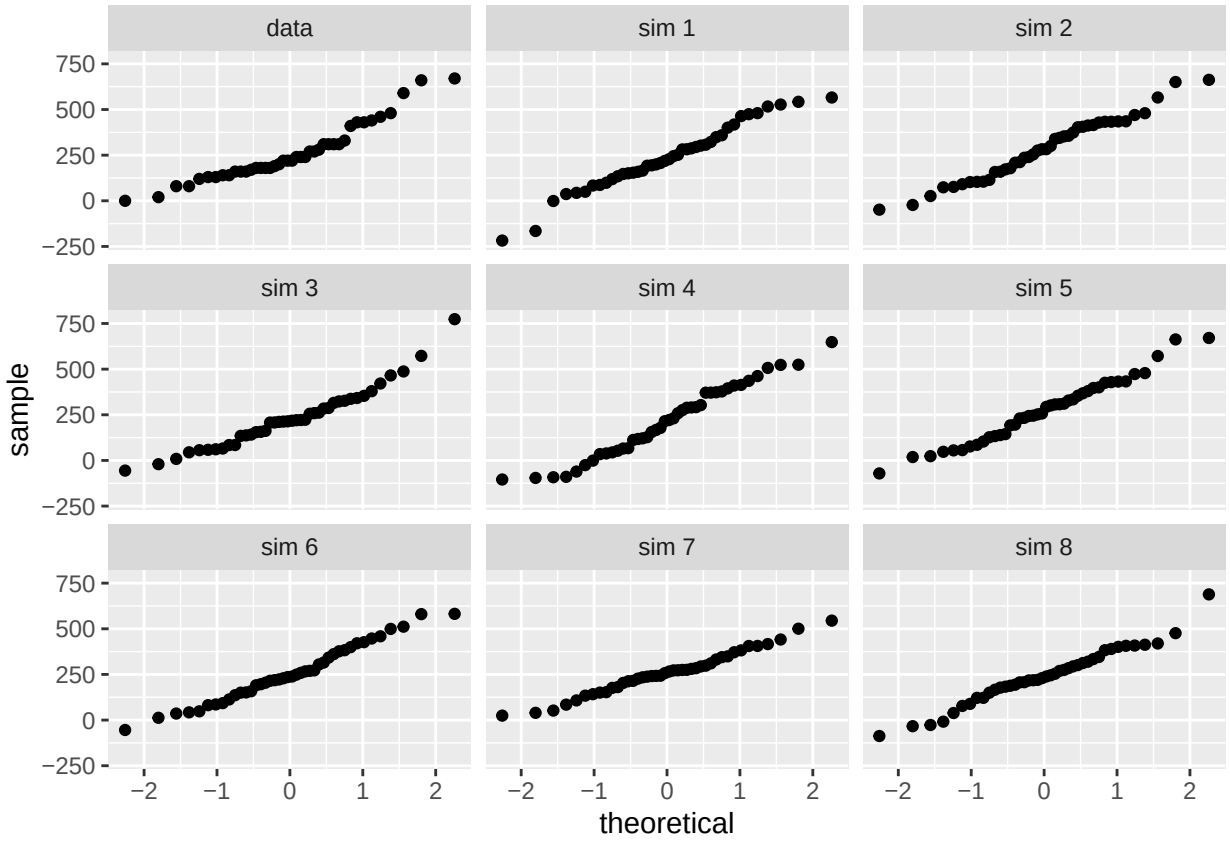
All the points don't fall on the line, however the trends of both plots are very similar. They both trend upwards, however the probability plot of the real data has a higher slope and thus ends past 600 while the simulated one is just past 400. Nonetheless, they're both pretty similar.

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?

The normal probability plot for the calories from fat does look similar to the plots created for the simulated data, which means that the calories are nearly normal.

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

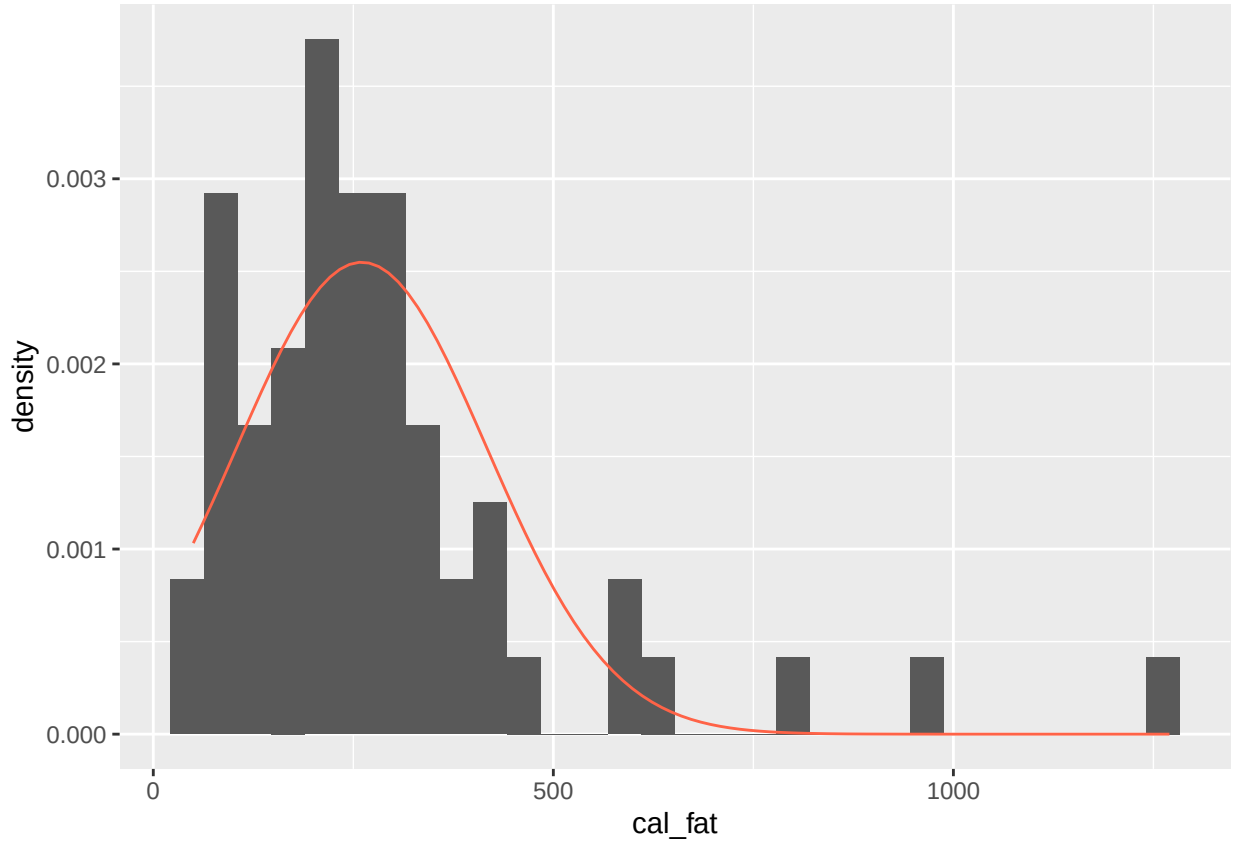


Exercise 5

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

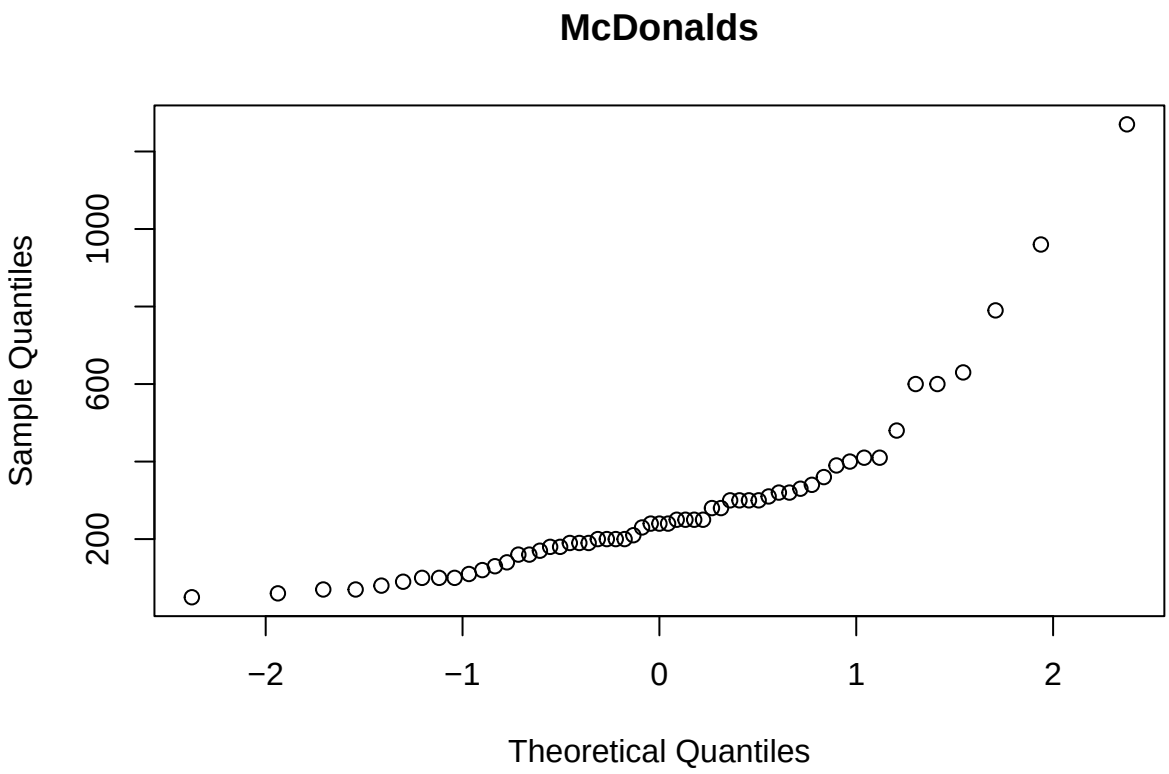
```
mdmean <- mean(mcdonalds$cal_fat)
mdsd   <- sd(mcdonalds$cal_fat)
ggplot(data = mcdonalds, aes(x = cal_fat)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = dqmean, sd = dqsd), col = "tomato")
```

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

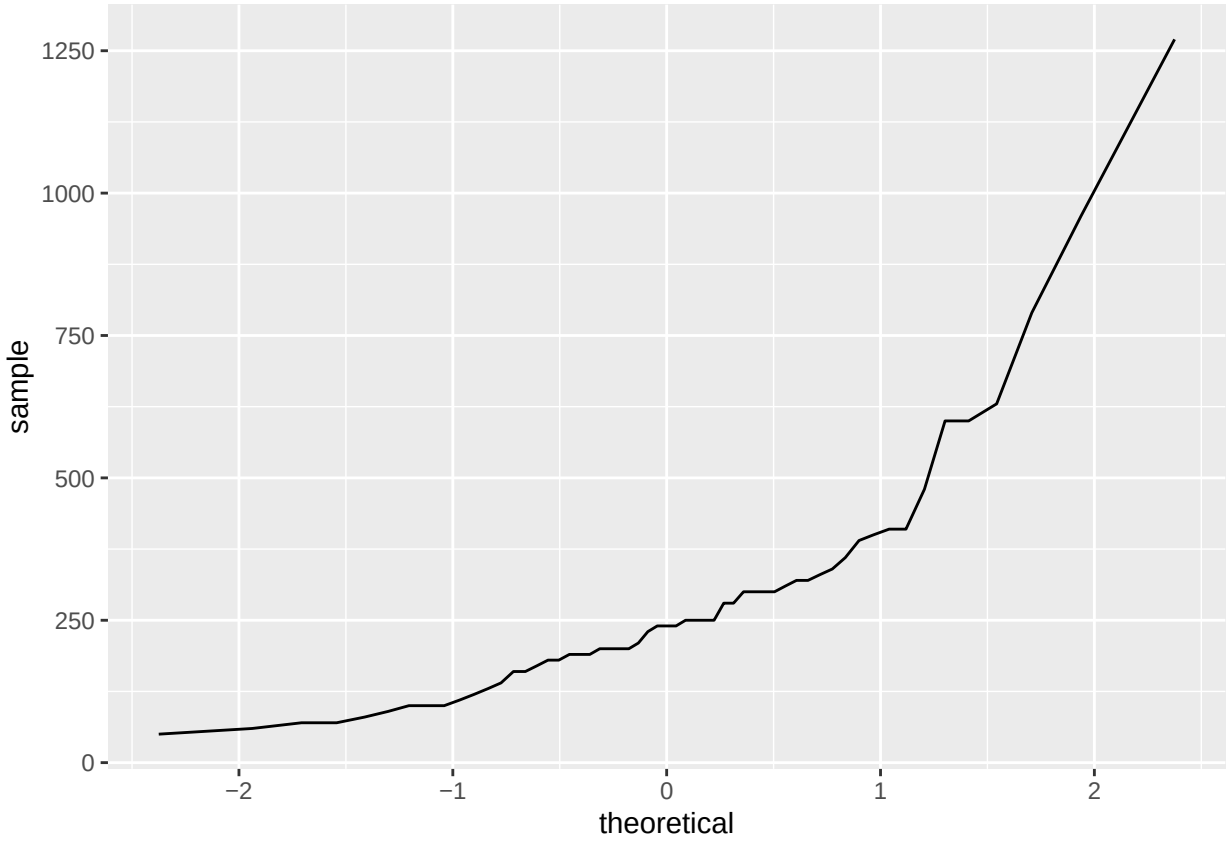


This actually doesn't look too bad in terms of following the curve, but let's continue onto the probability plots to really see if it's normal or not.

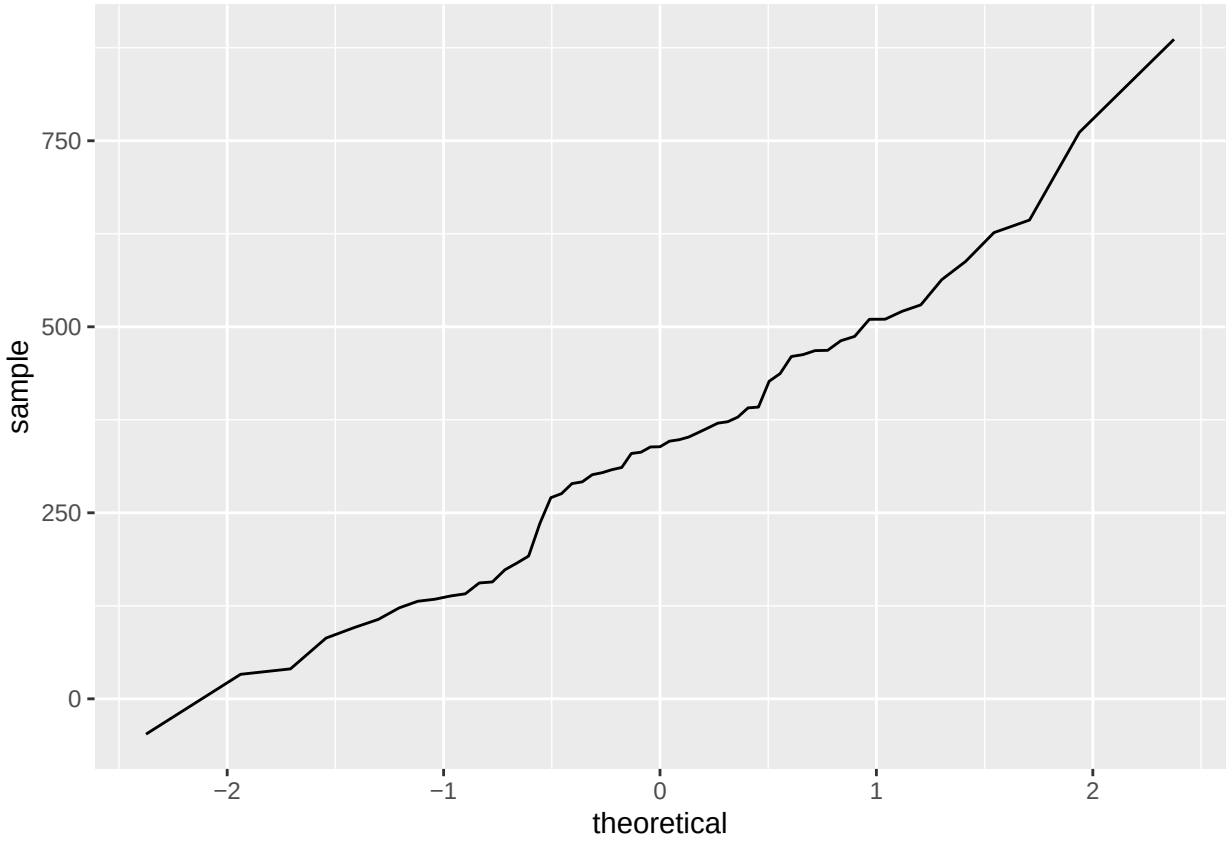
```
qqnorm(mcdonalds$cal_fat, main = "McDonalds")
```




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




```
md_sim_norm <- rnorm(n = nrow(mcdonalds), mean = mdmean, sd = mdsd)
ggplot(data = NULL, aes(sample = md_sim_norm)) +
  geom_line(stat = "qq")
```



These look pretty different when comparing the trends so I'm more likely to say that McDonald's data is not normal. The slopes are very different and the values are a lot more extreme in the actual data vs. the simulation. It also shows a right skew in the data based on the line shown.

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?

Question 1: What is the probability that a randomly chosen Burger King product has more than 500 calories from fat?

Theoretical:

```
burger_king <- fastfood %>%
  filter(restaurant == "Burger King")
bkmean <- mean(burger_king$cal_fat)
bksd <- sd(burger_king$cal_fat)
1 - pnorm(q = 500, mean = bkmean, sd = bksd)
```

```
## [1] 0.1963526
```

Empirical:

```
burger_king %>%
  filter(cal_fat > 500) %>%
  summarise(percent = n() / nrow(burger_king))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.171
```

0.1964 vs. 0.171 – these are relatively close which could imply that BK's data is relatively normal!

Question 2: What is the probability that a randomly chosen McDonald's product has more than 500 calories from fat?

Theoretical:

```
1 - pnorm(q = 500, mean = mdmean, sd = mdsd)
```

```
## [1] 0.165895
```

Empirical:

```
mcdonalds %>%
  filter(cal_fat > 500) %>%
  summarise(percent = n() / nrow(mcdonalds))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.105
```

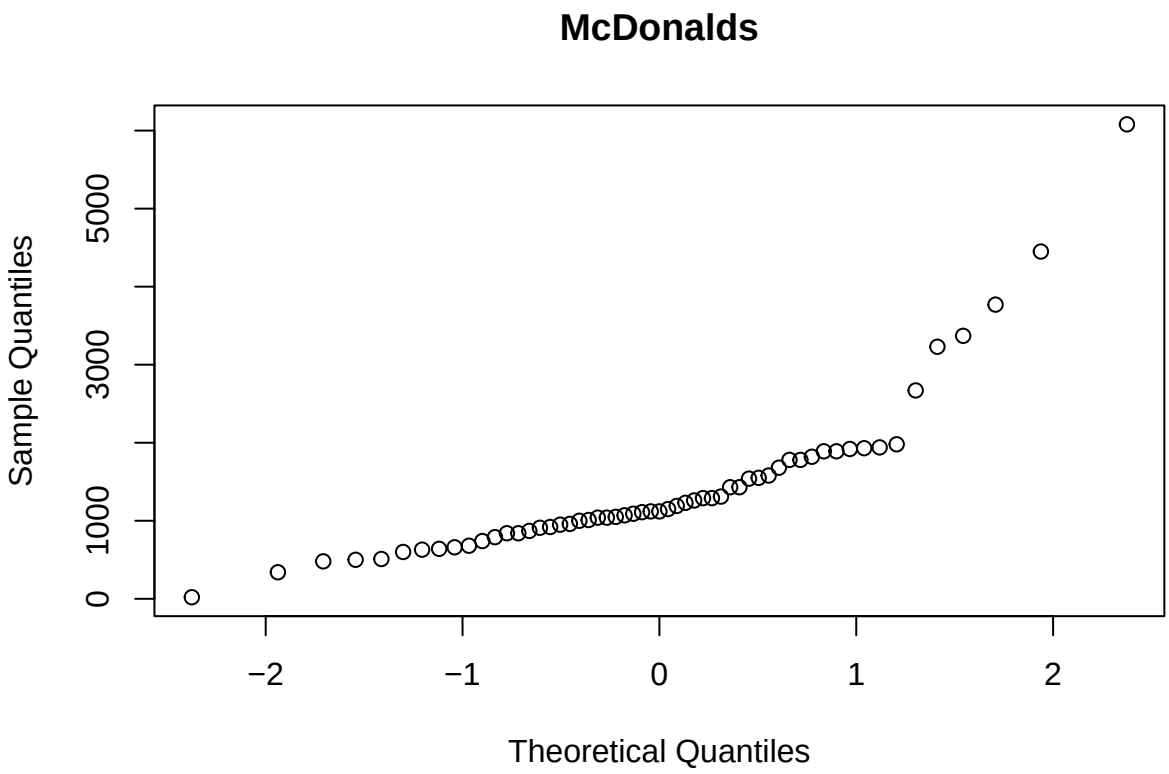
0.1659 vs. 0.105 – these aren't too close, but honestly not that bad considering McDonald's data isn't normally distributed.

Exercise 7

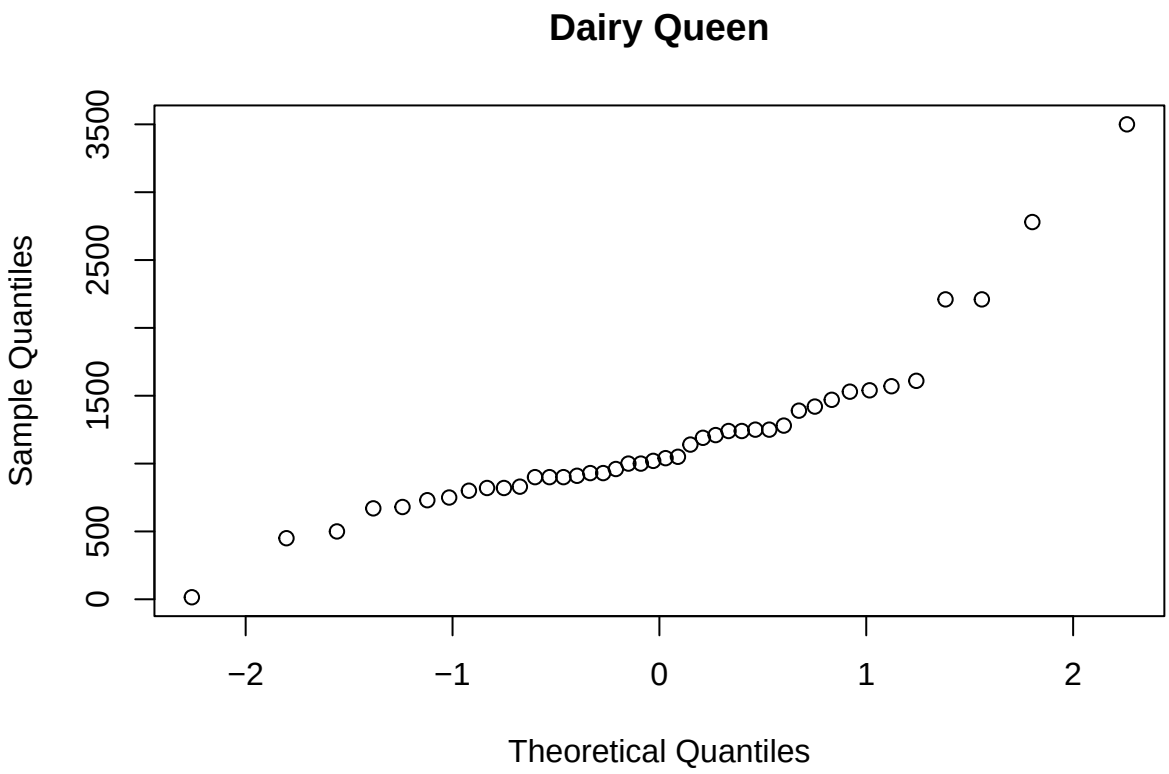
Now let's consider some of the other variables in the dataset. Out of all the different restaurants, which ones' distribution is the closest to normal for sodium?

```
arbys <- fastfood %>%
  filter(restaurant == "Arbys")
subway <- fastfood %>%
  filter(restaurant == "Subway")
sonic <- fastfood %>%
  filter(restaurant == "Sonic")
taco_bell <- fastfood %>%
  filter(restaurant == "Taco Bell")
cfa <- fastfood %>%
  filter(restaurant == "Chick Fil-A")

qqnorm(mcdonalds$sodium, main = "McDonalds")
```

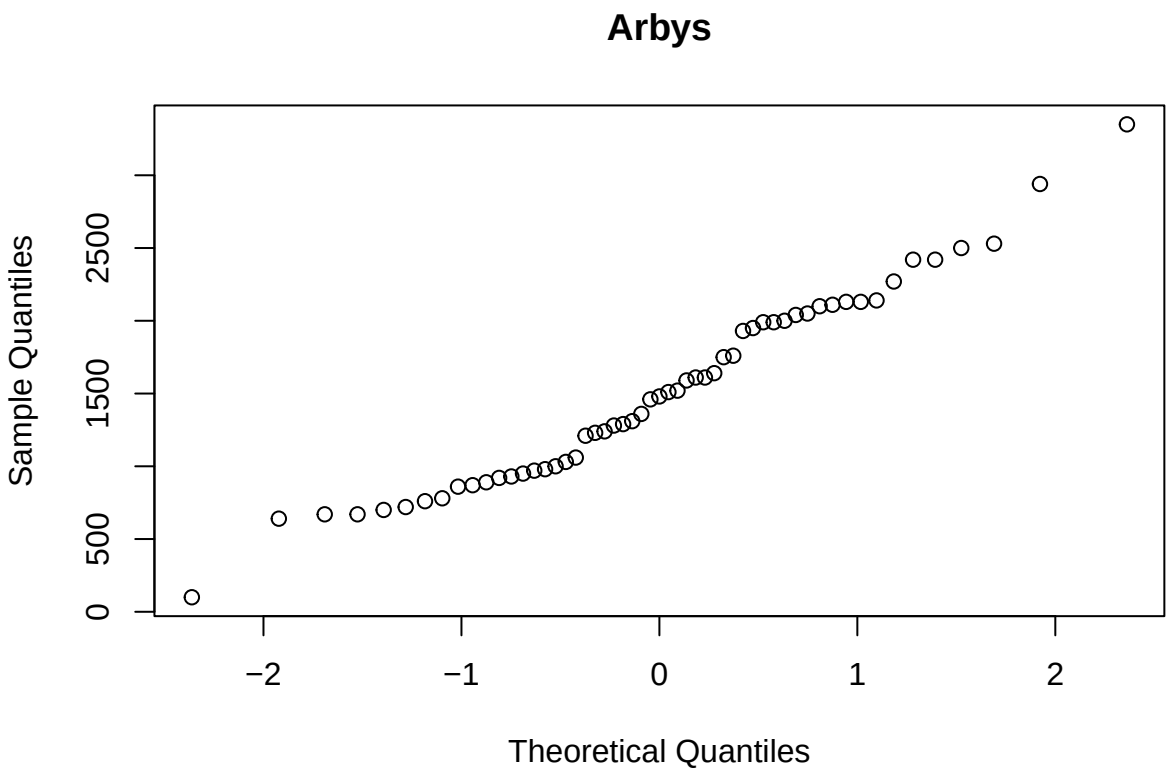



```
qqnorm(dairy_queen$sodium, main = "Dairy Queen")
```

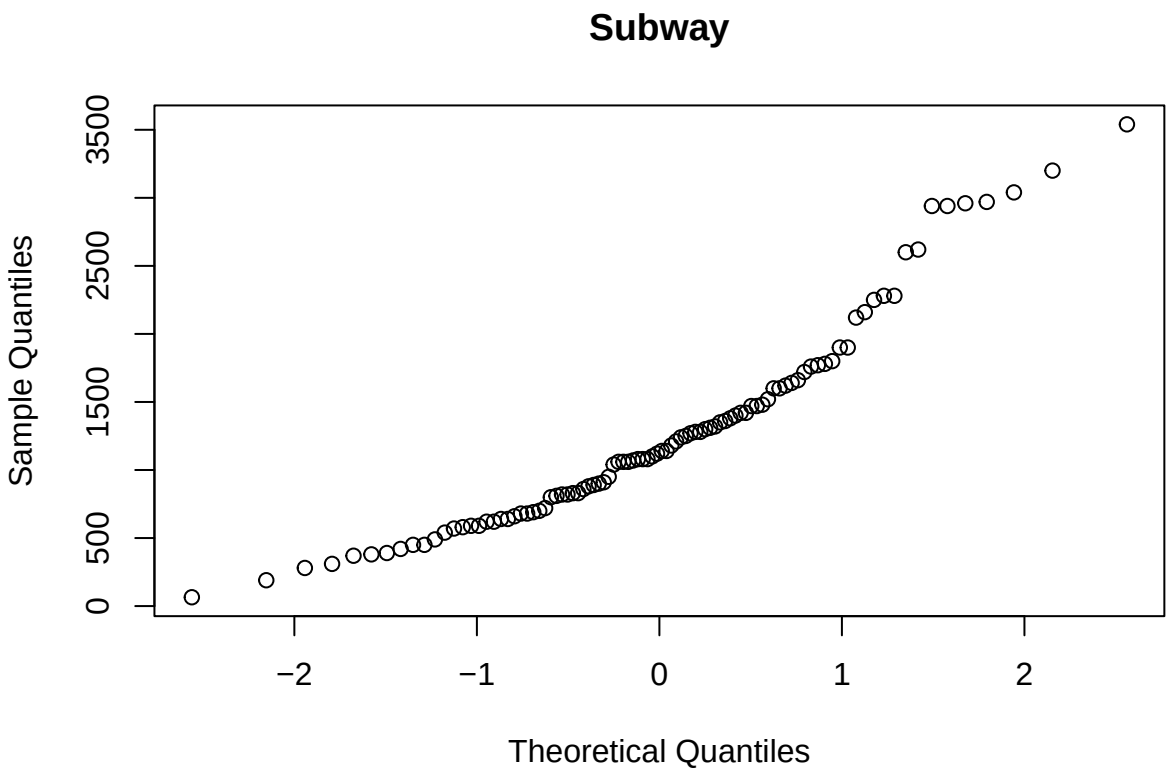



```
qqnorm(burger_king$sodium, main = "Burger King")
```

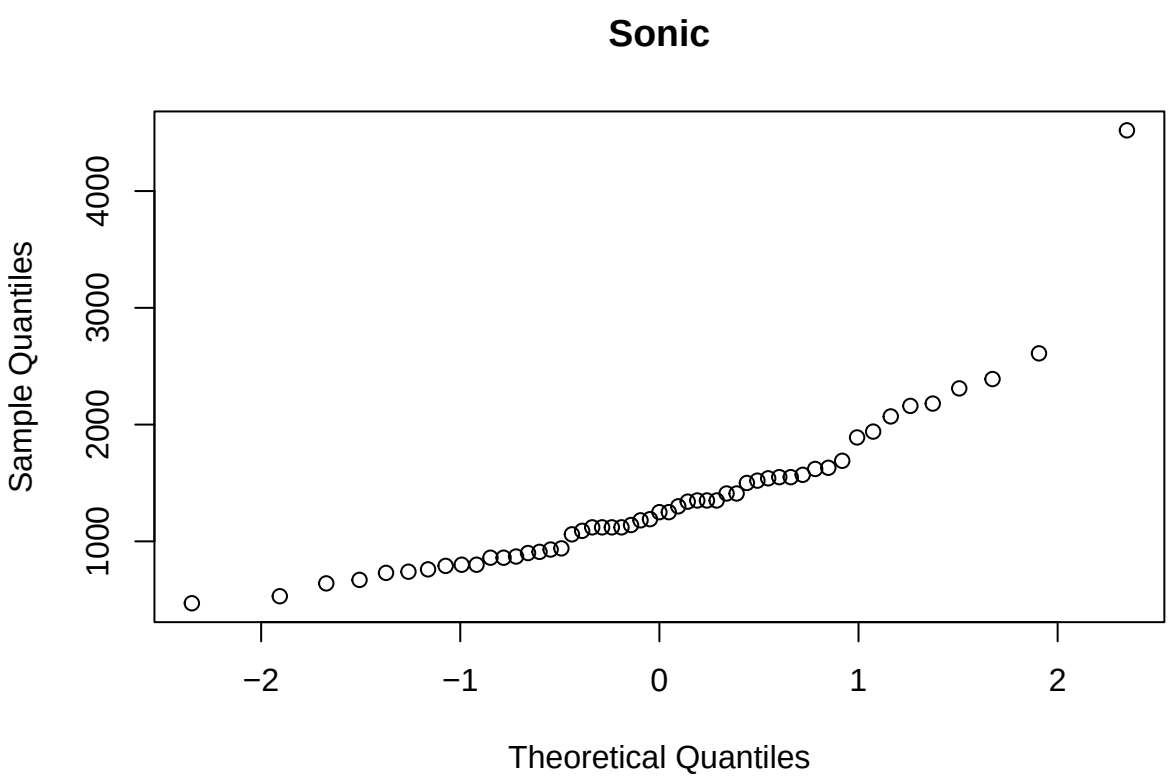

```
qqnorm(arbys$sodium, main = "Arbys")
```



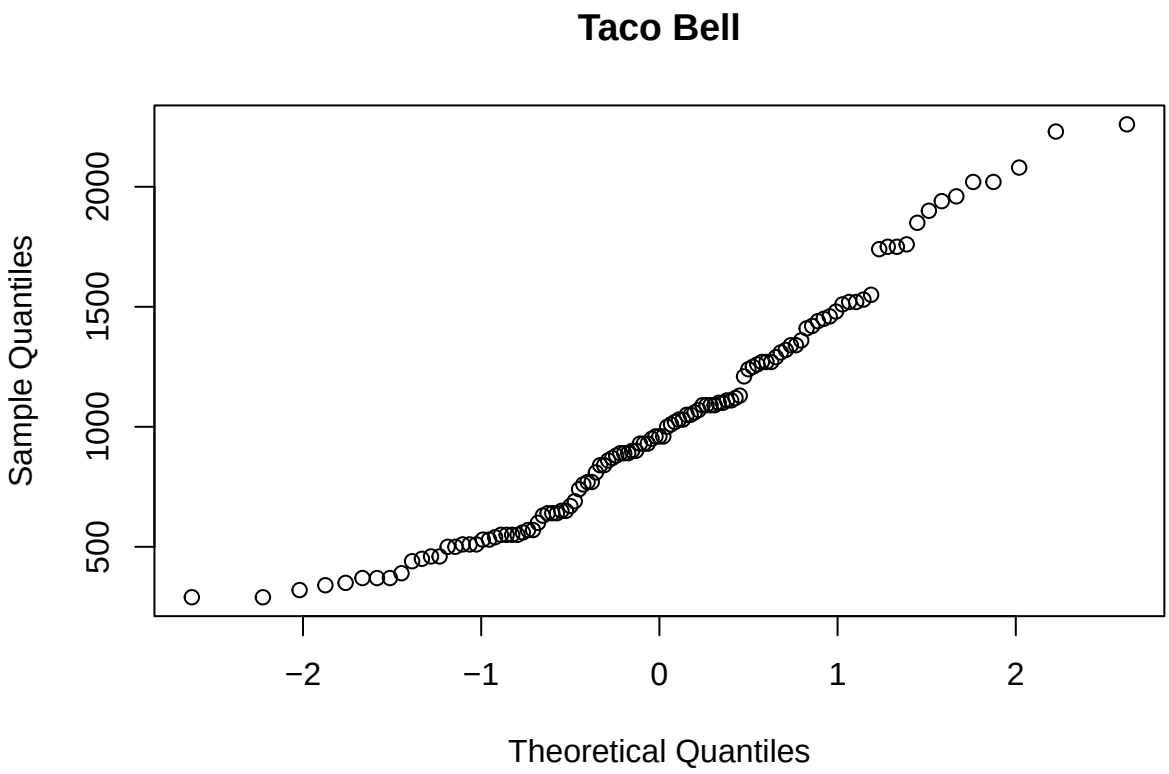

```
qqnorm(subway$sodium, main = "Subway")
```



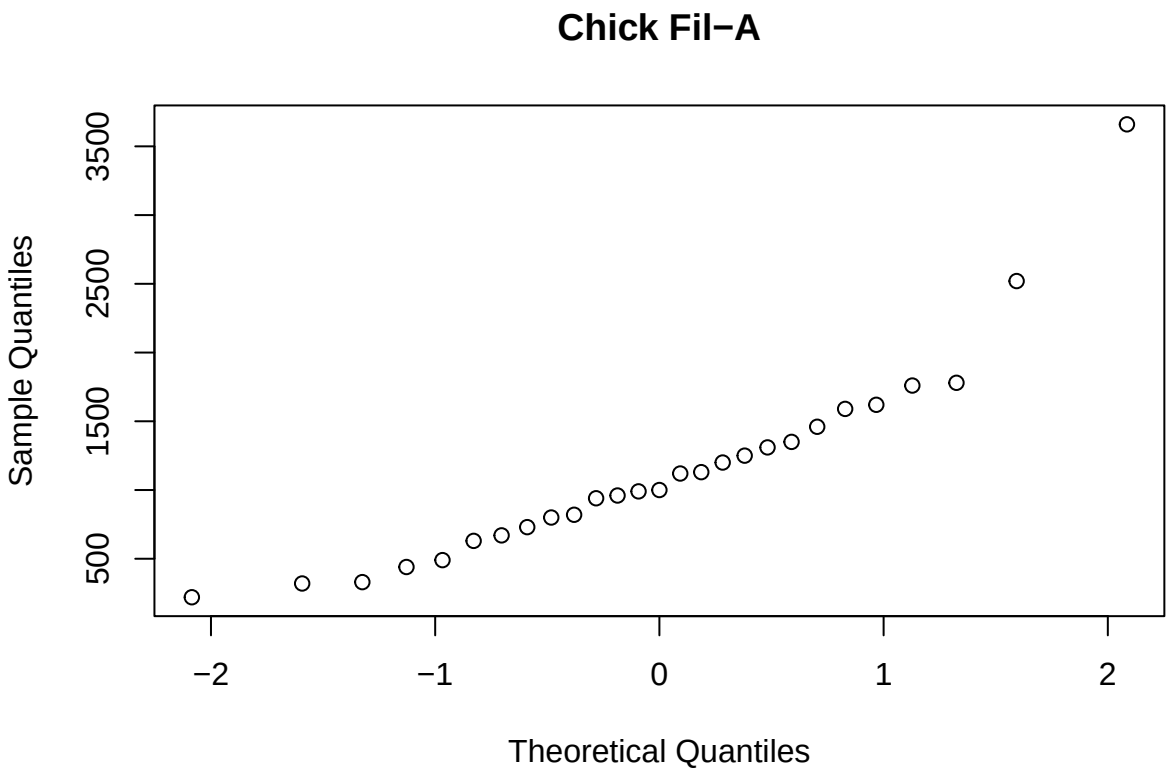

```
qqnorm(sonic$sodium, main = "Sonic")
```




```
qqnorm(taco_bell$sodium, main = "Taco Bell")
```




```
qqnorm(cfa$sodium, main = "Chick Fil-A")
```



I would say that Burger King's sodium levels are the most normal based on the charts shown above.

Exercise 8

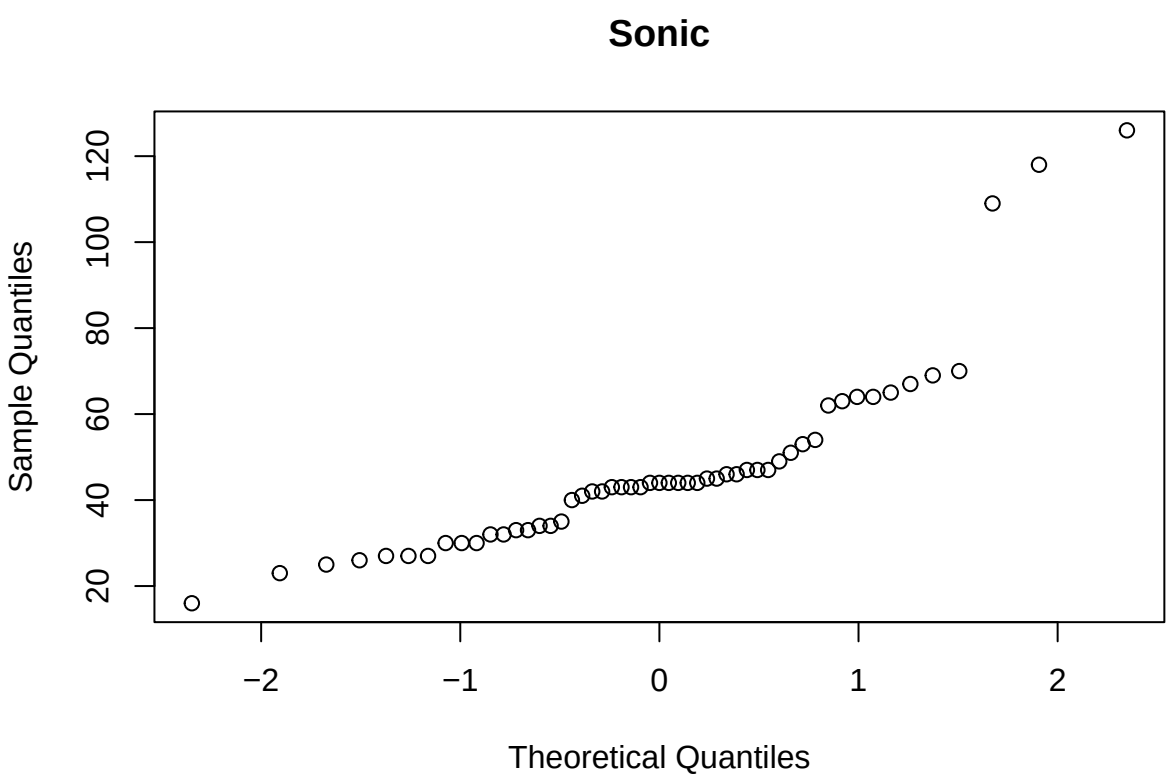
Note that some of the normal probability plots for sodium distributions seem to have a stepwise pattern. why do you think this might be the case?

Stepwise patterns usually occur in data with discrete values, however sodium levels are a continuous measure; the only reason I can think of that there are steps is because there are likely huge outliers for things with sodium in them, such as fries or burgers.

Exercise 9

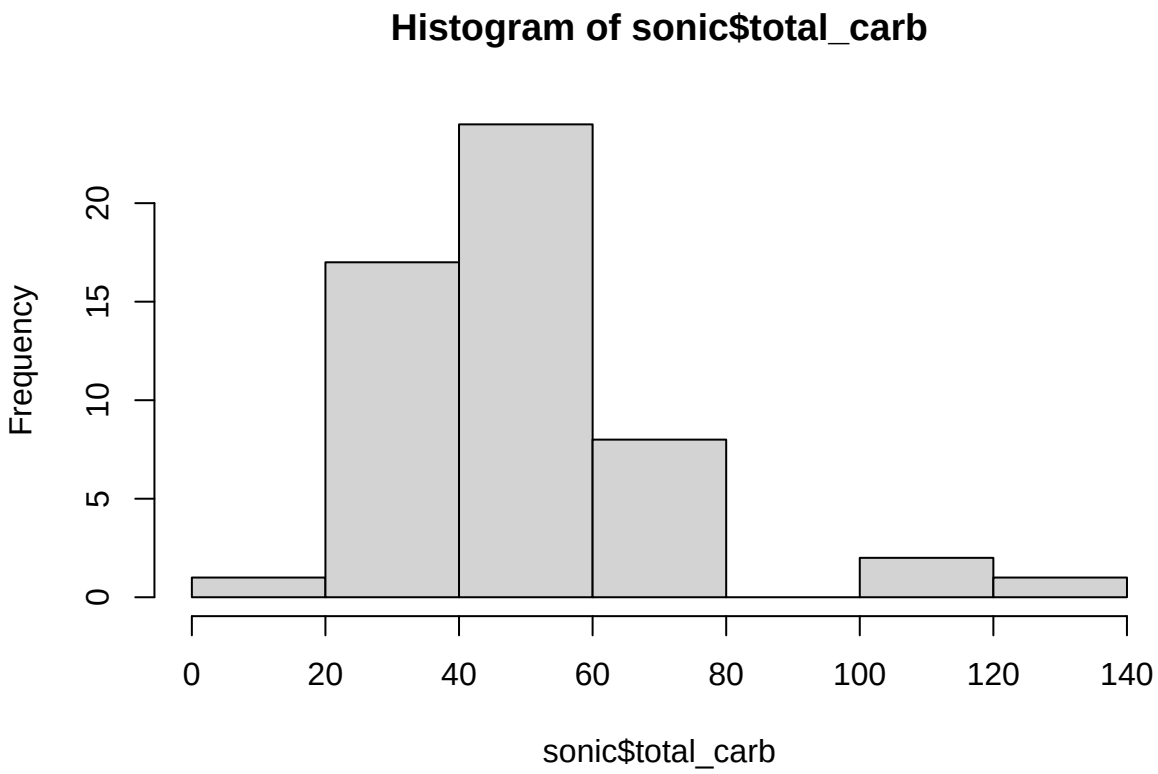
As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for the total carbohydrates from a restaurant of your choice. Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram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

```
qqnorm(sonic$total_carb, main = "Sonic")
```



For Sonic, I would say that this data is skewed towards the right due to there being a huge jump on the right side of the chart.

```
hist(sonic$total_carb)
```



Based on the histogram, the data does seem to be skewed to the right.