Data 606 - Lab 4

Cameron Smith

2020-09-27

```
library(tidyverse)
library(openintro)
library(cowplot)
```

Exercise 1

McDonalds has a median of 240, a mean of 285, and a max of 1270. Dairy Queen has a median of 220, a mean o 260, and a max of 670.

The fat calories in their offerings are fairly similar, though McDonalds has a significantly higher max - the 20 piece chicken tenders. The standard deviation for McDonalds is thus quite a bit higher as well (221 versus 156).

The distribution of each has been plotted below, and they are fairly similar - i.e. a unimodal, fairly normal distribution with some outliers (right skewed).

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

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

# View some quick summary statistics
summary(mcdonalds$cal_fat)
```

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

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

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

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

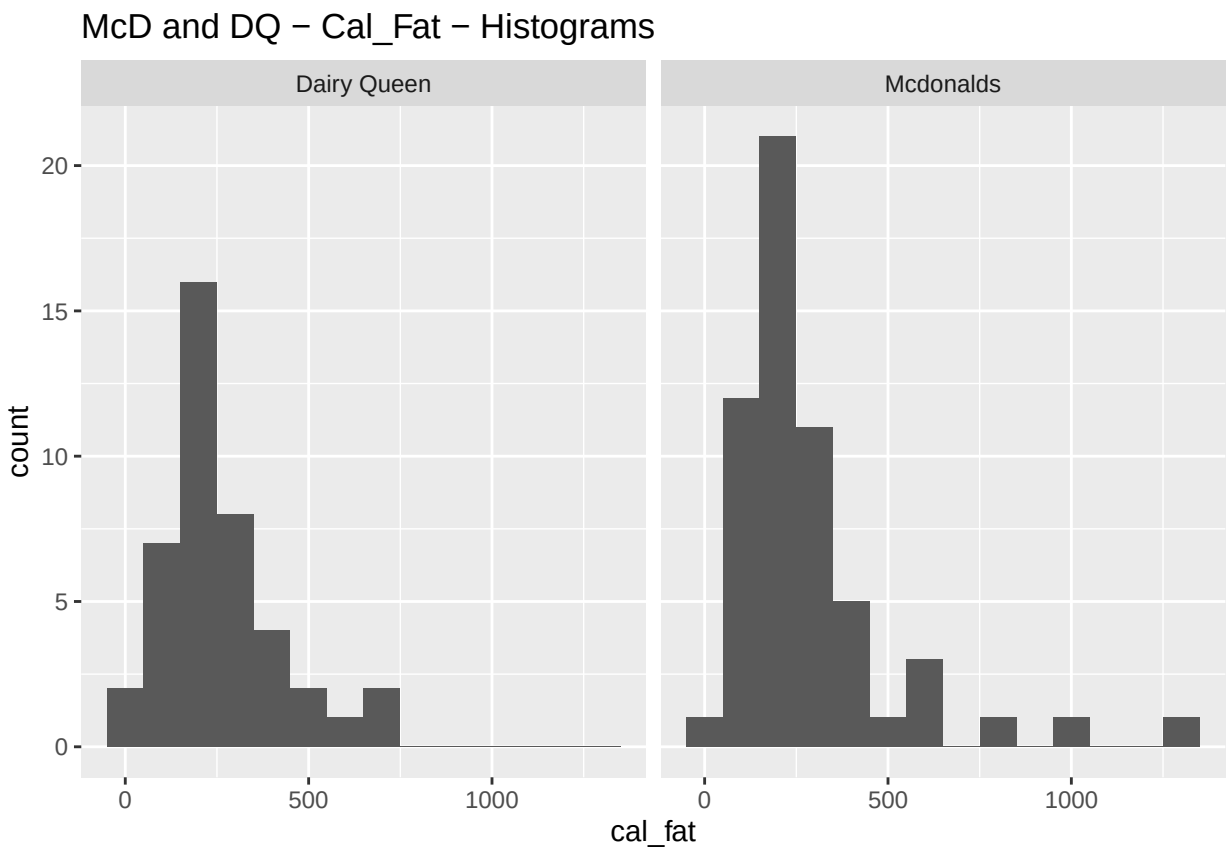
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   160.0   220.0   260.5   310.0   670.0
```

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

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

```
mc_queen <- rbind(mcdonalds, dairy_queen)

# Histograms showing both restaurants
ggplot(data = mc_queen, aes(x = cal_fat)) +
  geom_histogram(binwidth = 100) +
```

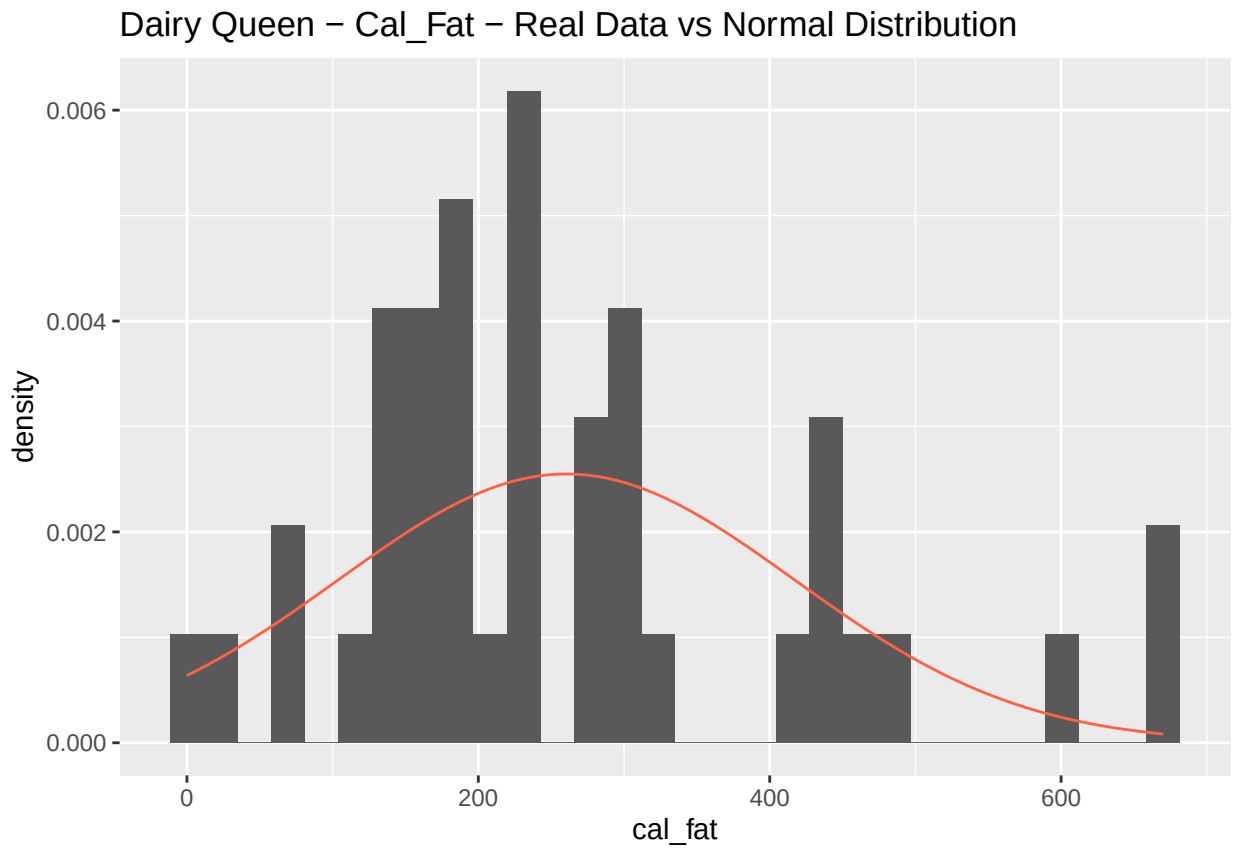
```
facet_wrap(~restaurant) +
ggtitle("McD and DQ - Cal_Fat - Histograms")
```



Exercise 2

Based on this plot, I would say the data follows a fairly normal distribution although there are definitely some outliers - in particular to the right where the distribution is skewed.

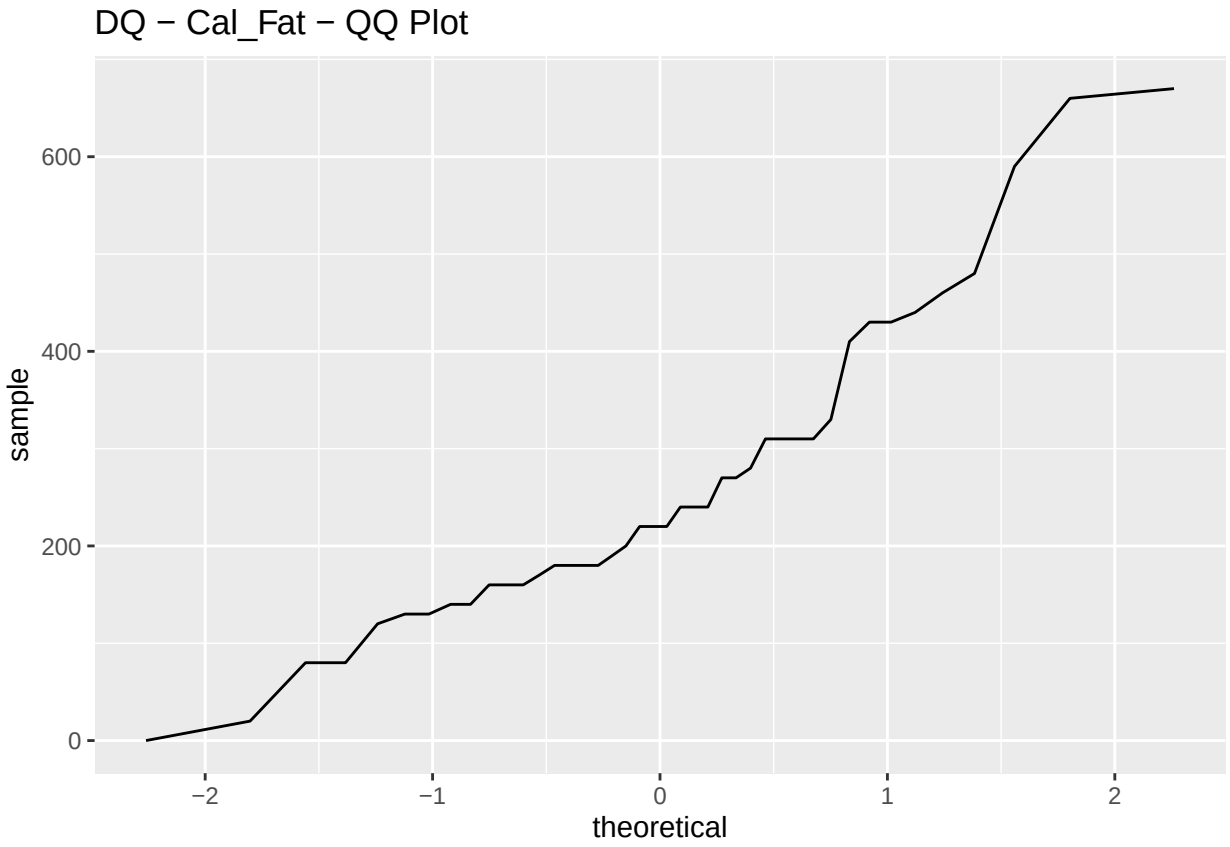
```
dqmean <- mean(dairy_queen$cal_fat)
dqsd   <- sd(dairy_queen$cal_fat)

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") +
        ggtitle("Dairy Queen - Cal_Fat - Real Data vs Normal Distribution")
```



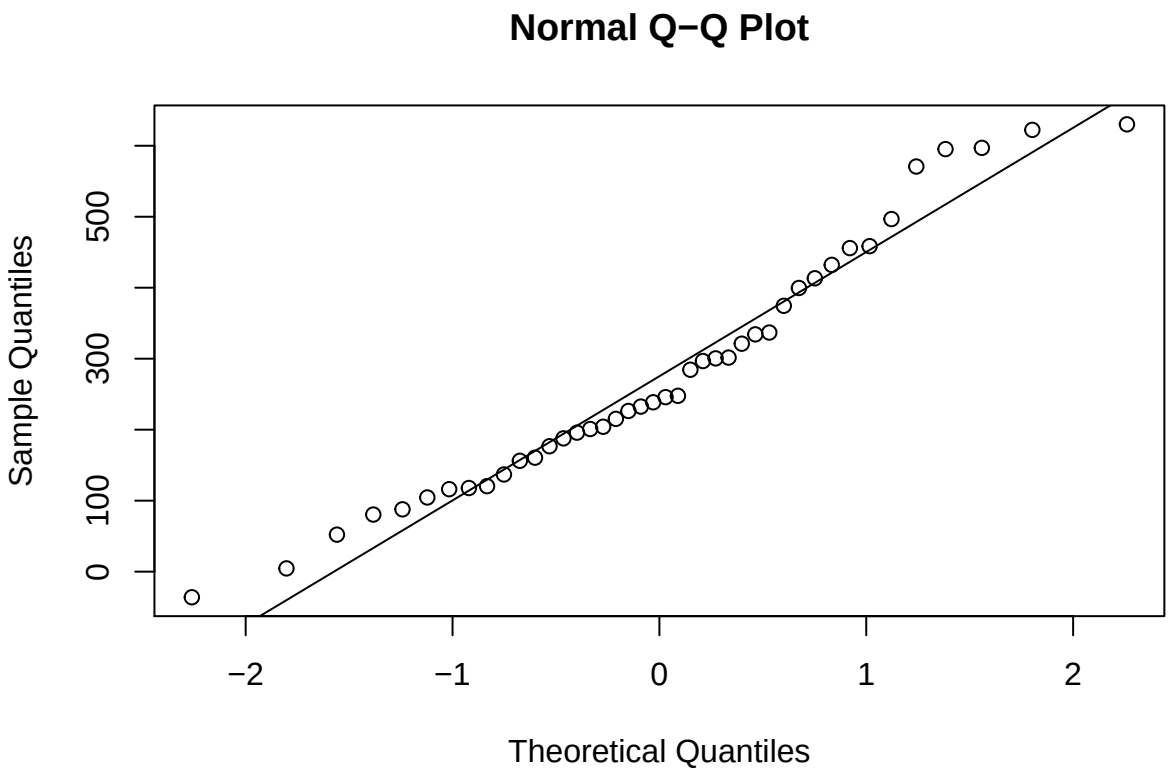
Exercise 3

Not all the points fall on the line within a QQ plot of sim_norm, though between -1 and 1 it is pretty close. The plot is pretty similar to that for the real data.

```
# Normal probability aka QQ plot for real data
ggplot(data = dairy_queen, aes(sample = cal_fat)) +
  geom_line(stat = "qq") +
  ggtitle("DQ - Cal_Fat - QQ Plot")
```




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

# QQ plot for simulated data.  Note - using qqnorm / qqline because I was unable to get ggplot to work
qqnorm(sim_norm); qqline(sim_norm)
```

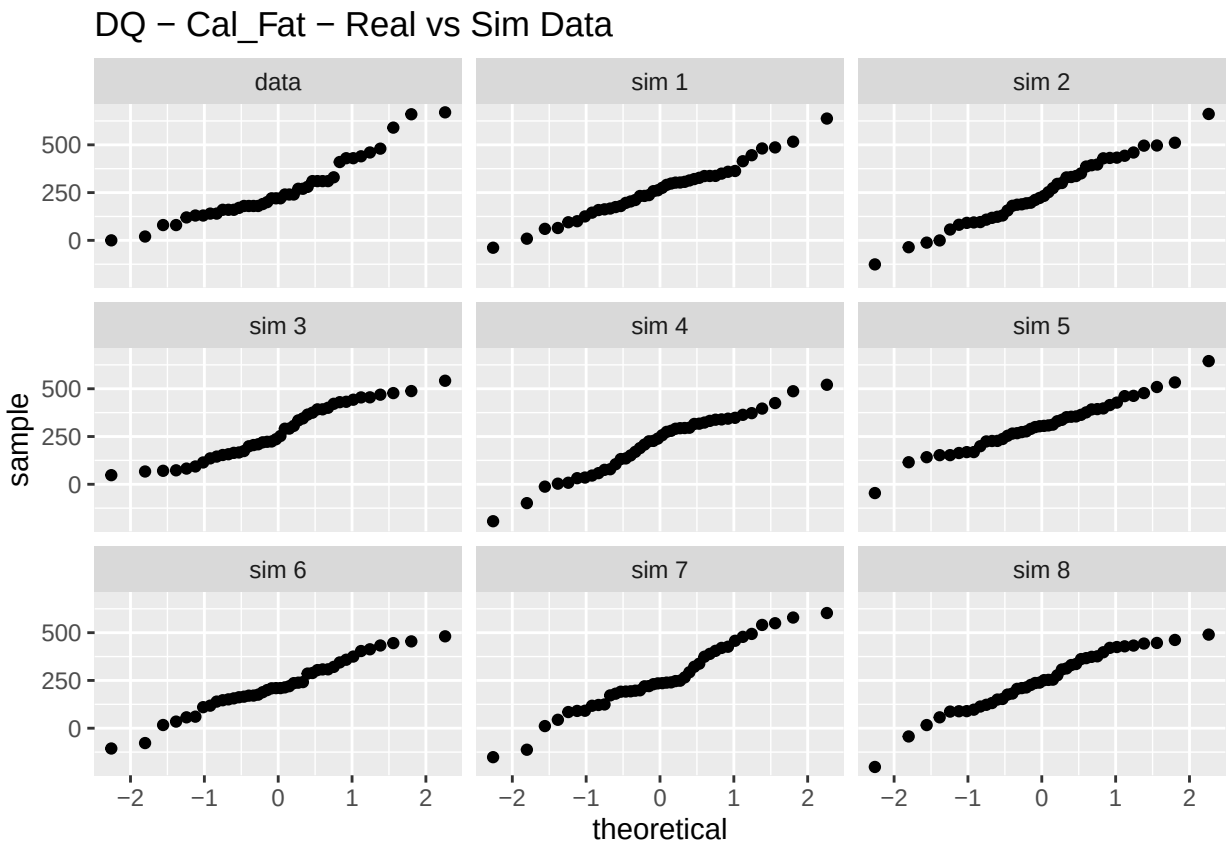



```
# Failed attempt with ggplot, excluding data argument as indicated in the instructions:
# ggplot(aes(sample = sim_norm)) +
# geom_line(stat = "qq")
```

Exercise 4

Yes, the plot from the real data looks pretty similar to the plots created for the simulated data. In particular both follow a normal distribution pretty closely from -1 to 1, and both have long tails at both ends.

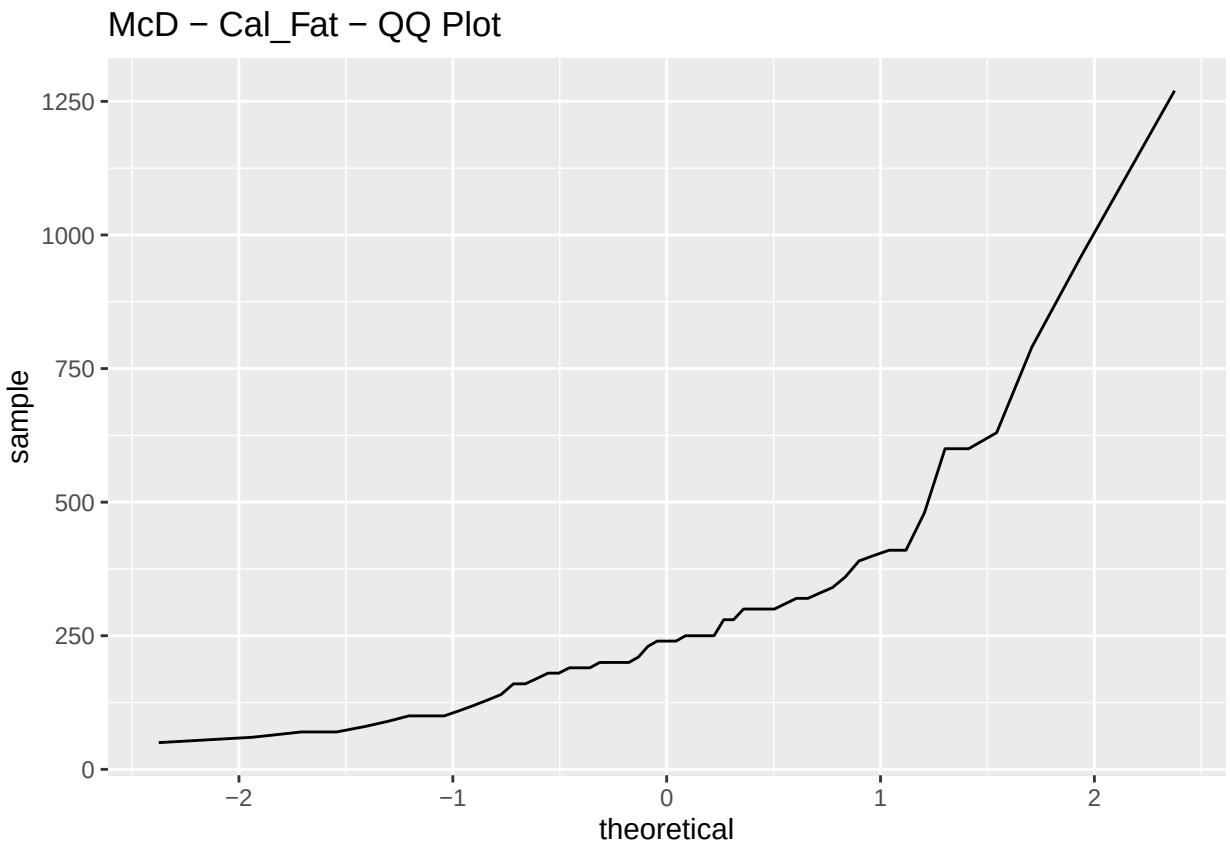
```
# Plots for simulated data
qqnormsim(sample = cal_fat, data = dairy_queen) +
  ggtitle("DQ - Cal_Fat - Real vs Sim Data")
```



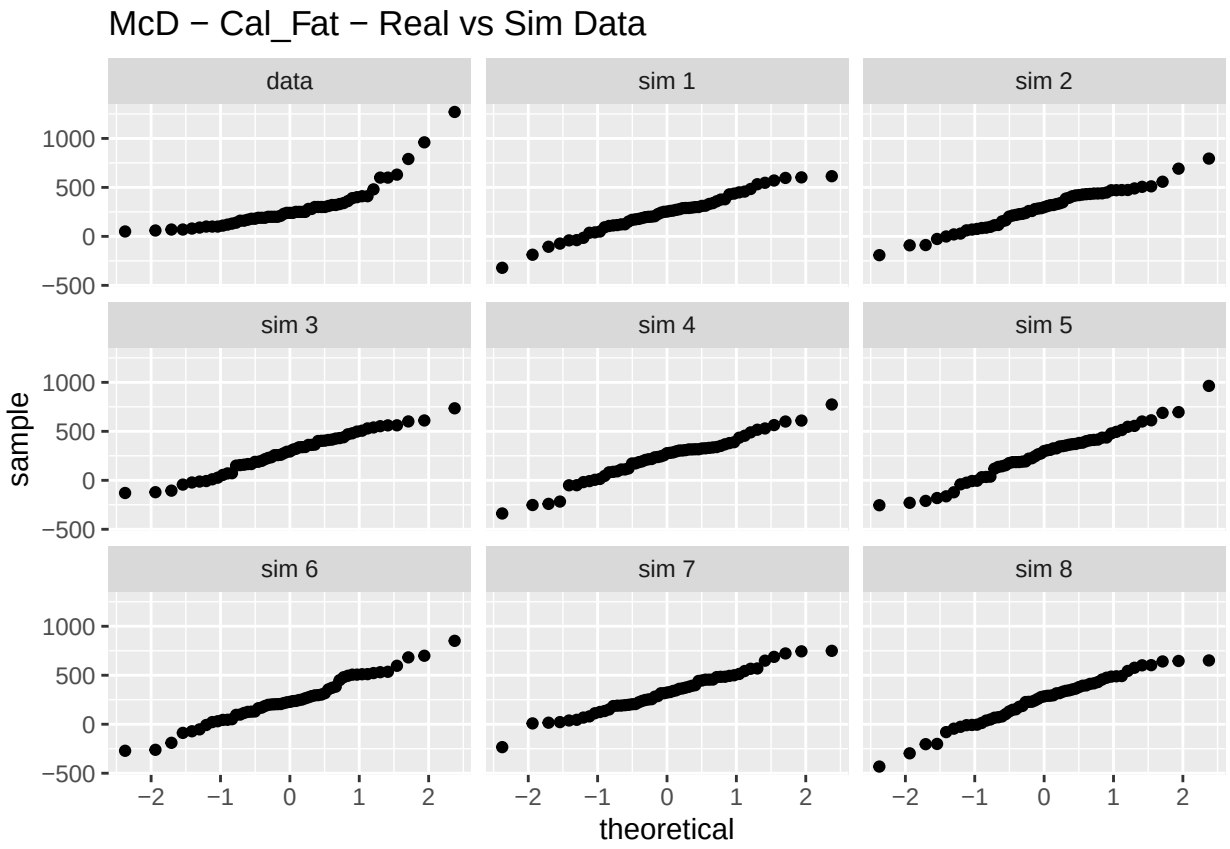
Exercise 5

Although the QQ plot appears to be quite different from that of Dairy Queen, in fact it does seem to resemble a normal distribution which we can see from comparing it with the sim data. It is significantly skewed at the top end though, likely due to a few outlier items with a disproportionate amount of fat calories. This can be visualized through a density plot, as seen below.

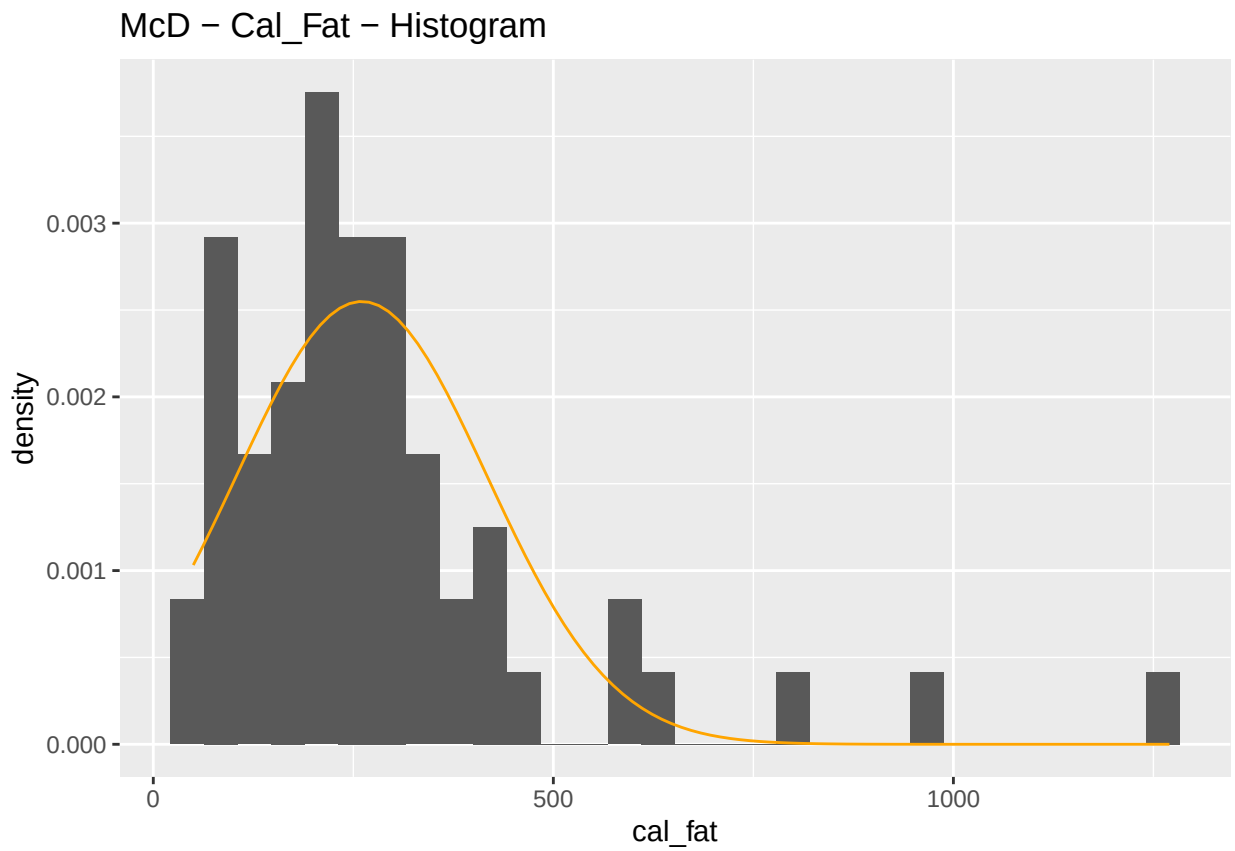
```
mcdmean <- mean(mcdonalds$cal_fat)
mcdsd   <- sd(mcdonalds$cal_fat)

ggplot(data = mcdonalds, aes(sample = cal_fat)) +
  geom_line(stat = "qq") +
  ggtitle("McD - Cal_Fat - QQ Plot")
```




```
sim_norm2 <- rnorm(n = nrow(mcdonalds), mean = mcdmean, sd = mcdsd)

# Multiple plots for simulated data
qqnormsim(sample = cal_fat, data = mcdonalds)+
  ggtitle("McD - Cal_Fat - Real vs Sim Data")
```




```
# Density plot based on real data
# Indicates a fairly normal distribution, heavily skewed to the right
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 = "orange")+
        ggtitle("McD - Cal_Fat - Histogram")
```


McD − Cal_Fat − Histogram
density
0.003
0.002
0.001
0.000
0
500
1000
cal_fat

Exercise 6

Probability questions I chose to answer: (1) What's the probability of an item on the McDonalds menu being under 600 fat_calories? (2) What's the probability of an item on the McDonalds menu having more than 20 grams of protein?

As detailed in the code and comments below, the McDonalds fat_calories data has a closer agreement - specifically a difference of less than 1% compared to 9% for the protein data.

```
# Calculate Z Score and probability for items > 600 fat_calories at Dairy Queen
1 - pnorm(q = 600, mean = dqmean, sd = dqsd)
```

```
## [1] 0.01501523
```

```
# Empirically:
# We see difference of approximately 3% from theoretical probability calculated above
dairy_queen %>%
  filter(cal_fat > 600) %>%
  summarise(percent = n() / nrow(dairy_queen))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1  0.0476
```

```
## What's the probability of an item on the McDonalds menu being under 600 calories?
# Theoretical calculation
1 - pnorm(q = 600, mean = mcdmean, sd = mcdsd)
```

```
## [1] 0.07733771
```

```
# Empirically:
# We see a difference of less than 1%, indicating that McDonalds data is closer to a
# normal distribution
mcdonalds %>%
  filter(cal_fat > 600) %>%
  summarise(percent = n() / nrow(mcdonalds))
```

```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1  0.0702
```

```
## What's the probability of an item on the McDonalds menu having more than 20 grams of # protein?

mcd_protein_mean <- mean(mcdonalds$protein)
mcd_protein_sd   <- sd(mcdonalds$protein)

# Theoretical calculation:
1 - pnorm(q = 20, mean = mcd_protein_mean, sd = mcd_protein_sd)
```

```
## [1] 0.754449
```

```
# Empirically:
# We see a difference of approximately 9%, indicating that the data does not follow a
# normal distribution as closely as the fat_cal data does.
mcdonalds %>%
  filter(protein > 20) %>%
  summarise(percent = n() / nrow(mcdonalds))
```

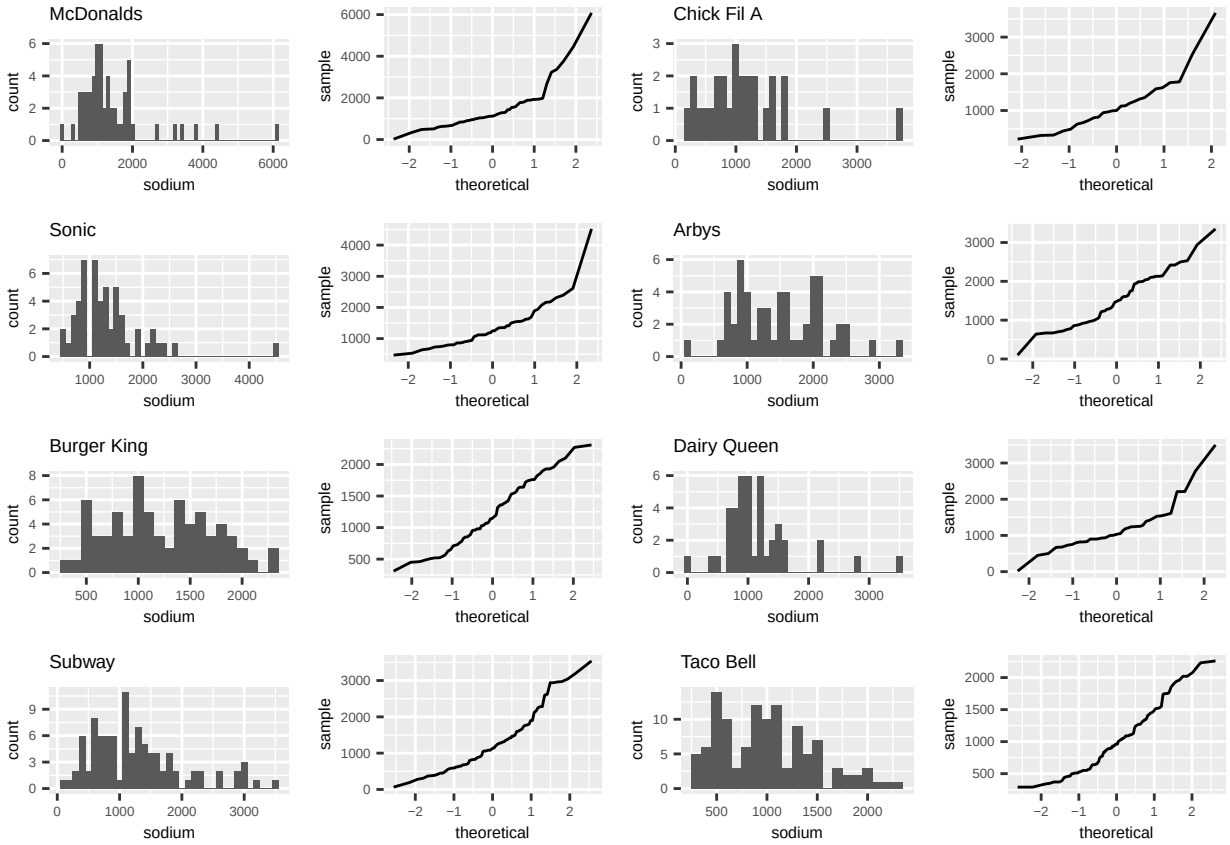
```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1   0.842
```

Exercise 7

It appears that Sonic and Subway are closest to having a normal distribution for sodium content.

```
# Side by side density and QQ plots per restaurant
# Note to self: There must be an easier way to do this in the future.
# Populate data frame of plots using for loop?

# McDonalds
plota1 <- ggplot(data = mcdonalds, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("McDonalds") +
  theme(text = element_text(size=6))
plota2 <- ggplot(data = mcdonalds, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Chick Fil A
chickfila <- fastfood %>%
  filter(restaurant=="Chick Fil-A")
plotb1 <- ggplot(data = chickfila, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Chick Fil A") +
  theme(text = element_text(size=6))
plotb2 <- ggplot(data = chickfila, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Sonic
sonic <- fastfood %>%
  filter(restaurant=="Sonic")
plotc1 <- ggplot(data = sonic, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Sonic") +
  theme(text = element_text(size=6))
plotc2 <- ggplot(data = sonic, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Arbys
arbys <- fastfood %>%
  filter(restaurant=="Arbys")
plotd1 <- ggplot(data = arbys, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Arbys") +
  theme(text = element_text(size=6))
plotd2 <- ggplot(data = arbys, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Burger King
burgerking <- fastfood %>%
  filter(restaurant=="Burger King")
plote1 <- ggplot(data = burgerking, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
```

```
  ggtitle("Burger King") +
  theme(text = element_text(size=6))
plote2 <- ggplot(data = burgerking, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Dairy Queen
plotf1 <- ggplot(data = dairy_queen, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Dairy Queen") +
  theme(text = element_text(size=6))
plotf2 <- ggplot(data = dairy_queen, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Subway
subway <- fastfood %>%
  filter(restaurant=="Subway")
plotg1 <- ggplot(data = subway, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Subway") +
  theme(text = element_text(size=6))
plotg2 <- ggplot(data = subway, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Taco Bell
tacobell <- fastfood %>%
  filter(restaurant=="Taco Bell")
ploth1 <- ggplot(data = tacobell, aes(x = sodium)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Taco Bell") +
  theme(text = element_text(size=6))
ploth2 <- ggplot(data = tacobell, aes(sample = sodium)) +
  geom_line(stat = "qq") +
  theme(text = element_text(size=6))

# Plot the charts side by side.
plot_grid(
  plota1, plota2, plotb1, plotb2, plotc1, plotc2, plotd1, plotd2, plote1, plote2, plotf1, plotf2, plotg
  ncol = 4)
```


McDonalds
count
sodium
sample
theoretical
Chick Fil A
Sonic
Arbys
Burger King
Dairy Queen
Subway
Taco Bell

Exercise 8

My best "educated" guess is that the stepwise patters are due to disproportionate counts of the variable at certain intervals. This could be for example due to restaurants aiming for a "sweet spot" of sodium content in this case - not too much, not too little - which has been incorporated into a large number of items on the menu.

Exercise 9

Based on the below QQ plot we can see that the Taco Bell data for total_carb is right skewed. This is further illustrated in the histogram.

```
ploti1 <- ggplot(data = tacobell, aes(x = total_carb)) +
  geom_histogram(binwidth = 10) +
  ggtitle("Taco Bell - Total_Carbs - Density Plot") +
  theme(text = element_text(size=10))

ploti2 <- ggplot(data = fastfood, aes(sample = total_carb)) +
  geom_line(stat = "qq") +
  ggtitle("Taco Bell - Total_Carbs - QQ Plot") +
  theme(text = element_text(size=10))

# Plot the charts side by side.
plot_grid(
  ploti1, ploti2)
```