431 Class 08

Thomas E. Love

2017-09-21

This Week

1. Comments on the Google Form & Assignment 1
2. Summaries withing subgroups (Notes: Ch 10)
3. Associations, Using Linear Models (Notes: Ch 11)

Comments on The Google Form (re: Assignment 1)

On a scale from 1 (extremely difficult) to 7 (extremely easy), how difficult did you find Assignment 1 to be? *

	1	2	3	4	5	6	7	
Extremely difficult	○	○	○	○	○	○	○	Extremely easy

What was frustrating to you in doing Assignment 1?

If it was not frustrating to you in any way, feel free to skip this question.

Long answer text

About how long (in minutes) did it take you to do Assignment 1? *

Comments on The Google Form (final item)

What would you like to be able to do by the end of 431 that you cannot do now?

This need not be anything related to what we've discussed so far.

More to come.

R Setup

```
library(viridis); library(gridExtra); library(ggridges)
library(knitr); library(pander)
library(tidyverse)

source("Love-boost.R")

nyfs1 <- read.csv("nyfs1.csv") %>% tbl_df
names(nyfs1)
```

```
[1] "subject.id"        "sex"
[3] "age.exam"          "bmi"
[5] "bmi.cat"           "waist.circ"
[7] "triceps.skinfold"
```

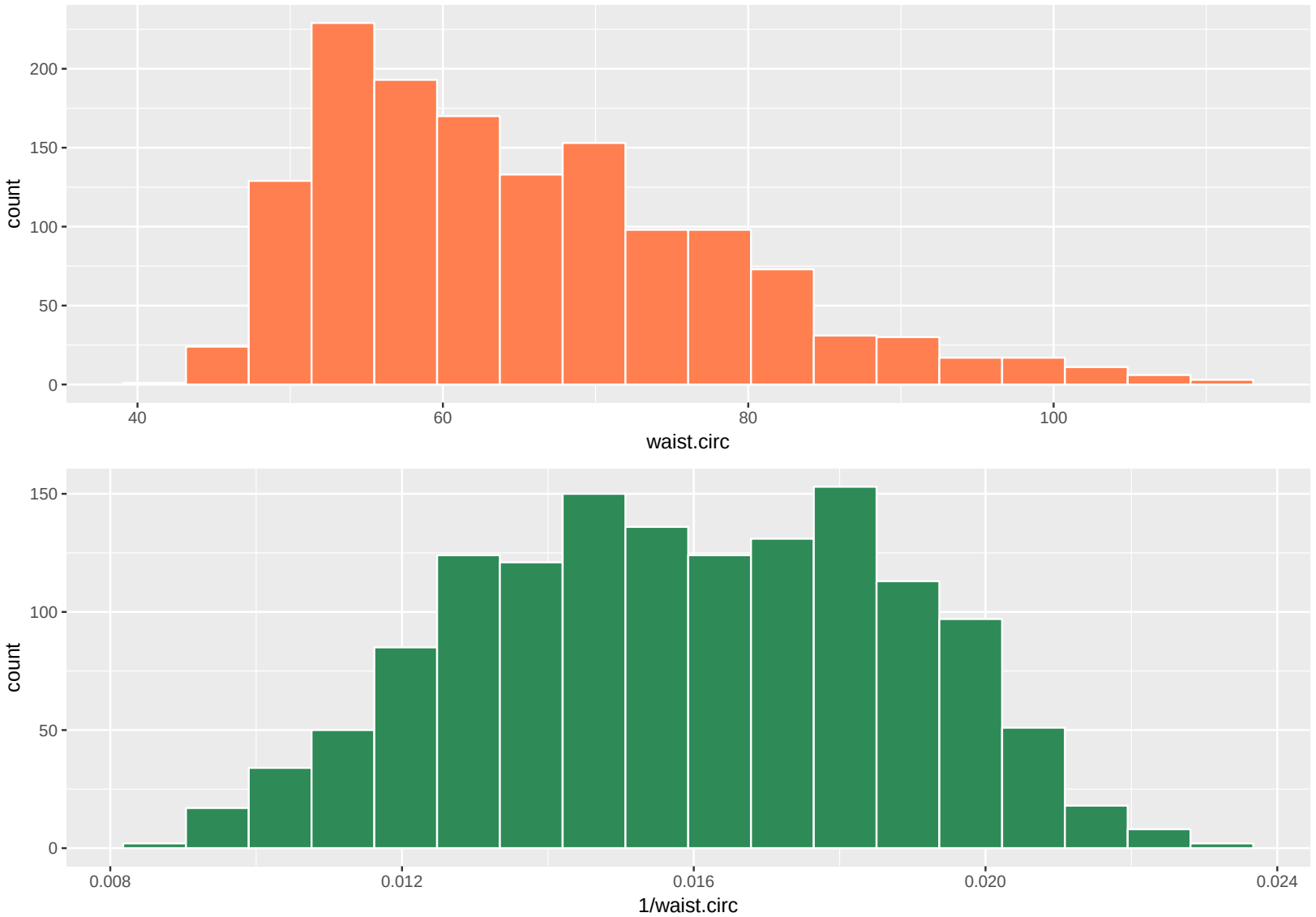
Transforming the Waist Circumference Data

```
p1 <- ggplot(nyfs1,
             aes(x = waist.circ)) +
  geom_histogram(bins = 18,
                 fill = "coral", col = "white")

p2 <- ggplot(nyfs1,
             aes(x = 1/waist.circ)) +
  geom_histogram(bins = 18,
                 fill = "seagreen", col = "white")

gridExtra::grid.arrange(p1, p2)
```

The Resulting Plot Array



Why Transform?

When we have unimodal but skewed data, we will often **transform** the data using a log, inverse, square root, square, etc. in order to obtain a new distribution which is closer to the Normal.

1. Sometimes we do this to facilitate comparisons.
    - Example: t-test to compare mean waist circumference among male children to female children
    - t-test requires that the distribution of the outcome in each sex be approximately Normal

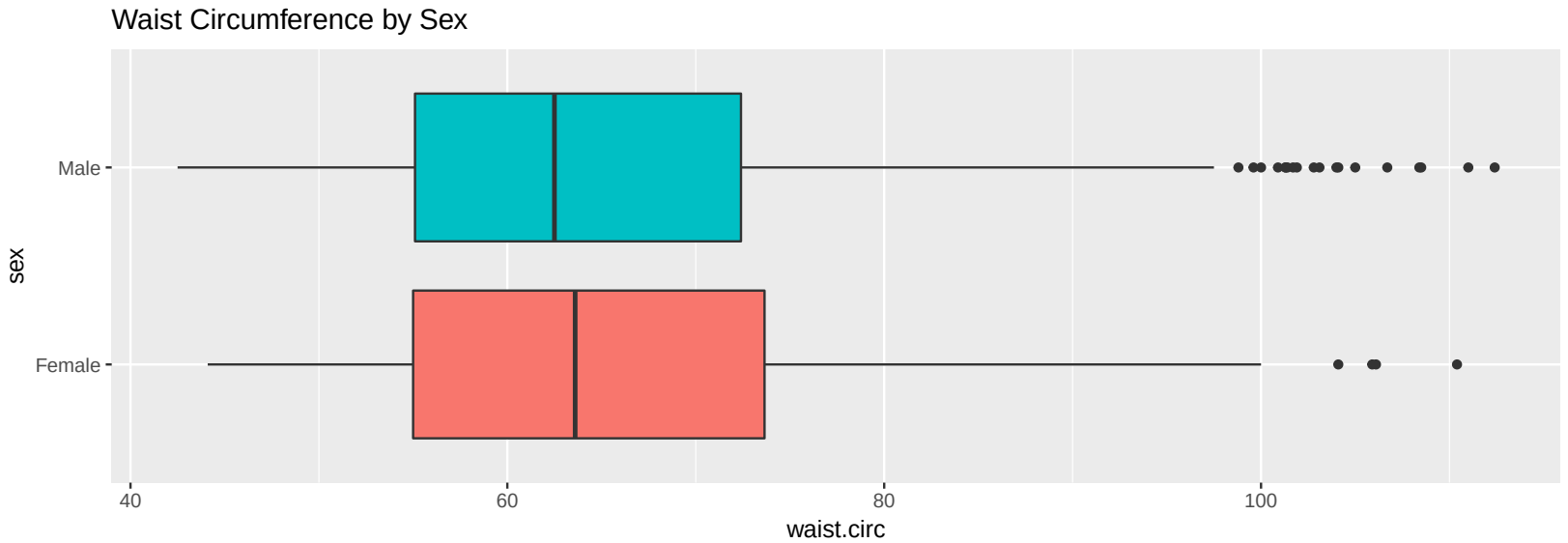
Does a Transformation Help with Comparison?

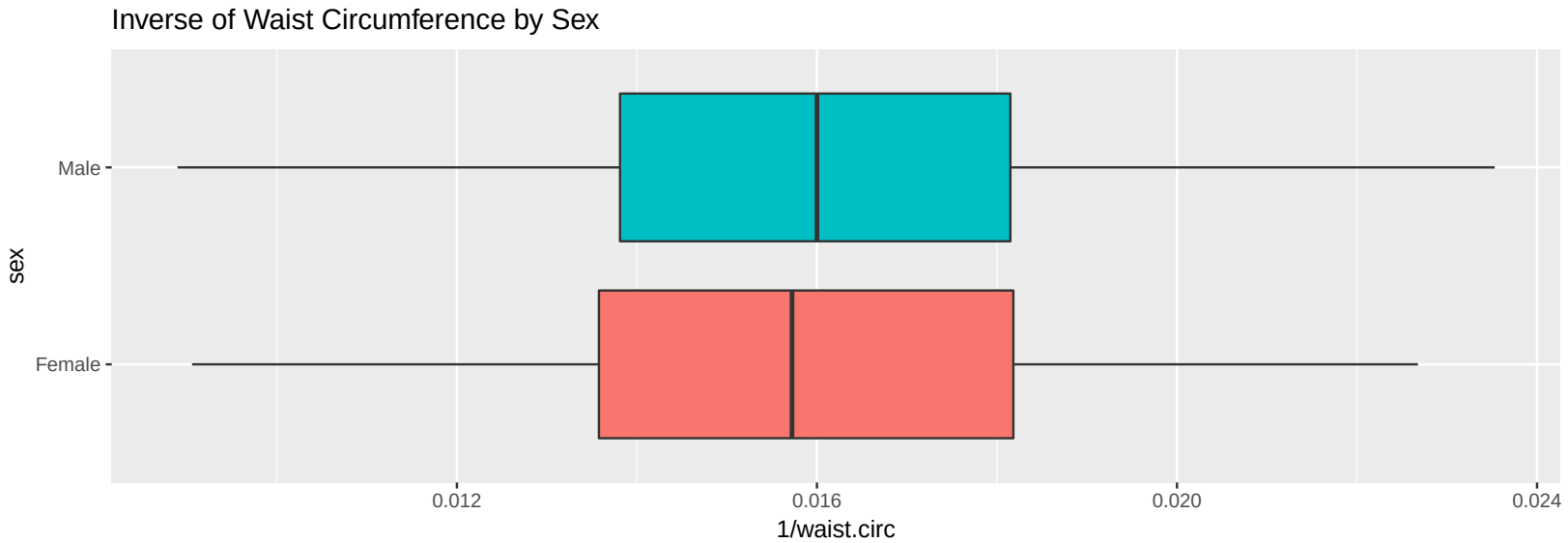
```
p1 <- ggplot(nyfs1, aes(x = sex, y = waist.circ,
                        fill = sex)) +
  geom_boxplot() +
  coord_flip() +
  guides(fill = FALSE) +
  labs(title = "Waist Circumference by Sex")

p2 <- ggplot(nyfs1, aes(x = sex, y = 1/waist.circ,
                        fill = sex)) +
  geom_boxplot() +
  coord_flip() +
  guides(fill = FALSE) +
  labs(title = "Inverse of Waist Circumference by Sex")

gridExtra::grid.arrange(p1, p2)
```

Boxplots of Waist Circumference by Sex





Why Transform?

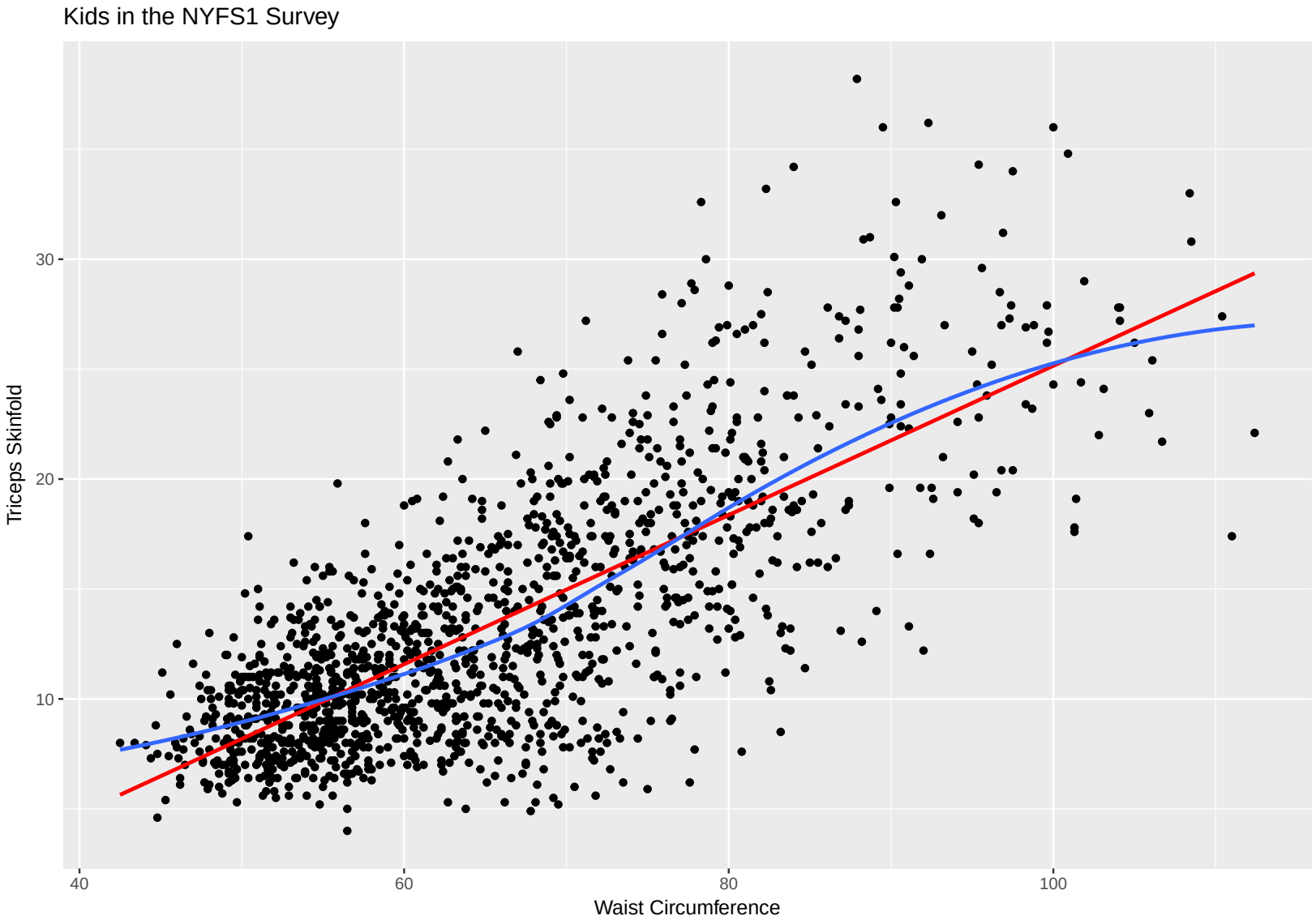
When we have unimodal but skewed data, we will often **transform** the data using a log, inverse, square root, square, etc. in order to obtain a new distribution which is closer to the Normal.

2. Sometimes we do this to facilitate model-building.
    - What is the association of waist circumference with triceps skinfold?
    - Transformations that "normalize" the distributions of skewed variables also can "linearize" an apparent association.

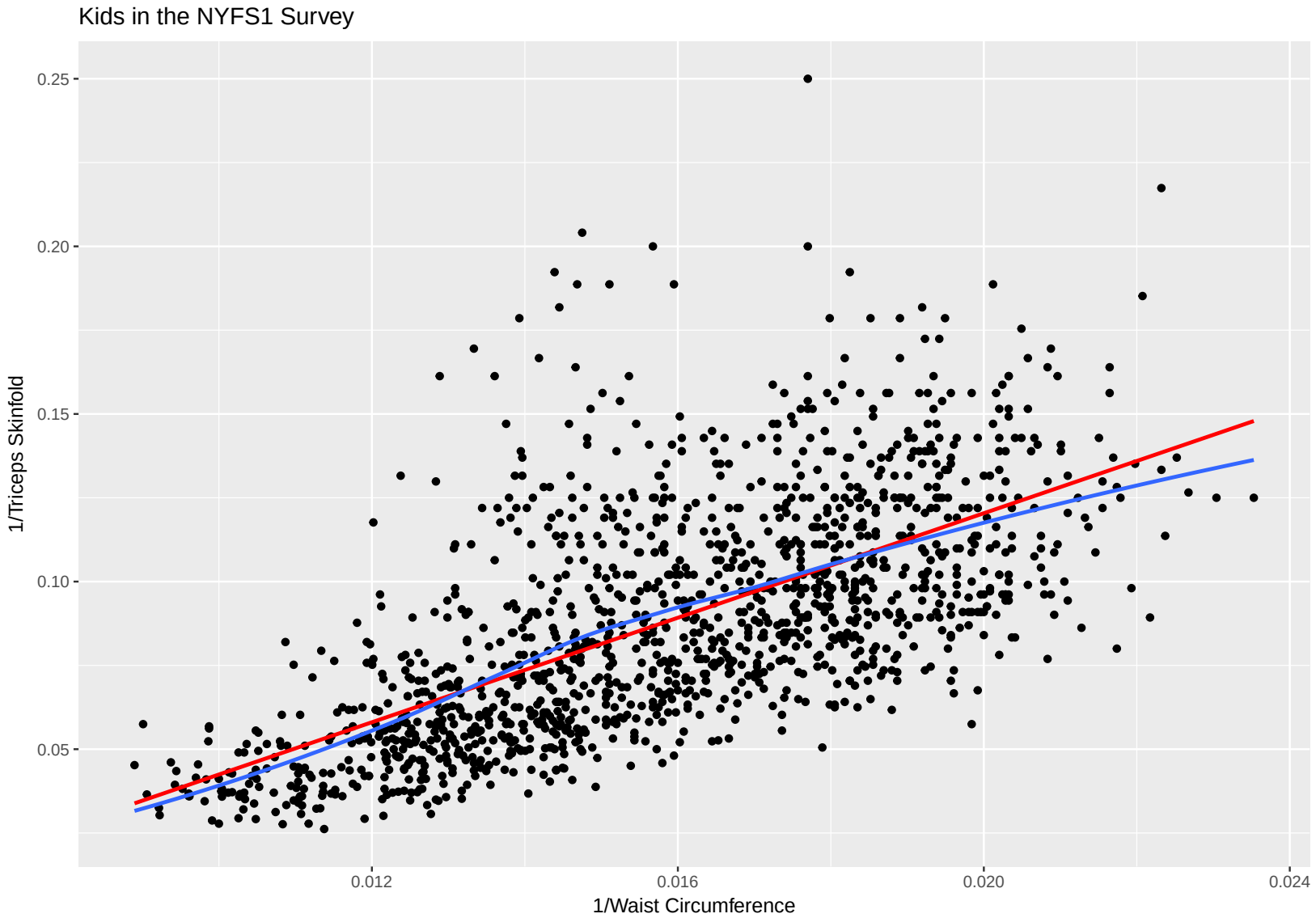
Is Waist Circumference related to Triceps Skinfold?

```
ggplot(nyfs1, aes(x = waist.circ, y = triceps.skinfold)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Waist Circumference",
       y = "Triceps Skinfold",
       title = "Kids in the NYFS1 Survey")
```

Is Waist Circumference related to Triceps Skinfold?



After Inverse Transformations (no real help here)



BMI categories

```
nyfs1 %>%
  count(bmi.cat)
```

```
# A tibble: 4 x 2
          bmi.cat     n
           <fctr> <int>
1   1 Underweight    42
2 2 Normal weight   926
3    3 Overweight   237
4           4 Obese   211
```

Use group_by and summarize together

```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2))
```

```
# A tibble: 4 x 5
          bmi.cat  mean median    sd skew1
           <fctr> <dbl>  <dbl> <dbl> <dbl>
1   1 Underweight  54.9   53.9   7.6  0.14
2 2 Normal weight  61.0   59.2   9.1  0.19
3    3 Overweight  71.1   72.0  11.8 -0.08
4         4 Obese  79.9   79.9  15.0  0.00
```

Using `knitr::kable` to present the table

```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2)) %>%
  knitr::kable()
```

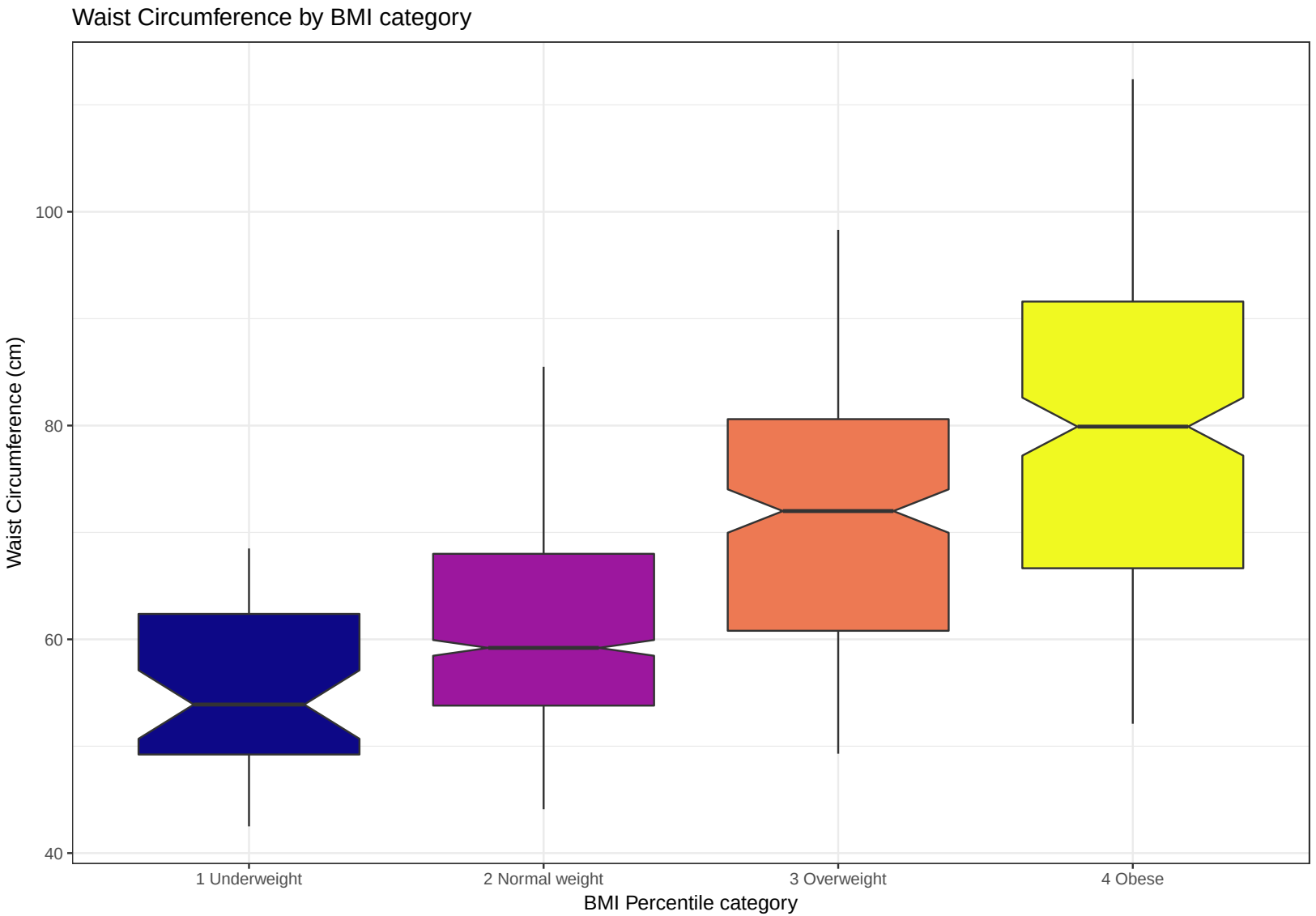
bmi.cat	mean	median	sd	skew1
1 Underweight	54.9	53.9	7.6	0.14
2 Normal weight	61.0	59.2	9.1	0.19
3 Overweight	71.1	72.0	11.8	-0.08
4 Obese	79.9	79.9	15.0	0.00

Using pander::pander to present the table

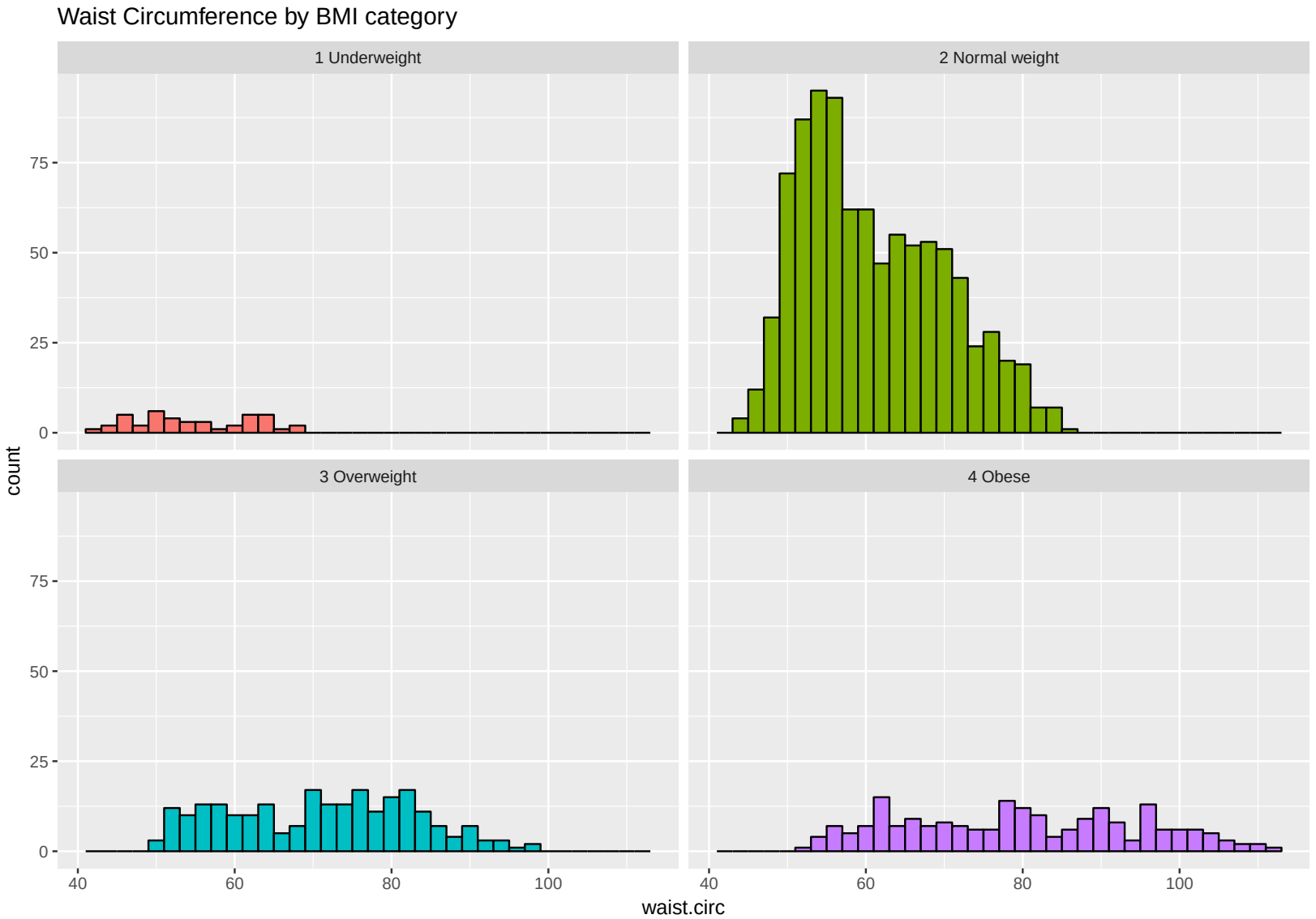
```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2)) %>%
  pander::pander()
```

bmi.cat	mean	median	sd	skew1
1 Underweight	54.9	53.9	7.6	0.14
2 Normal weight	61	59.2	9.1	0.19
3 Overweight	71.1	72	11.8	-0.08
4 Obese	79.9	79.9	15	0

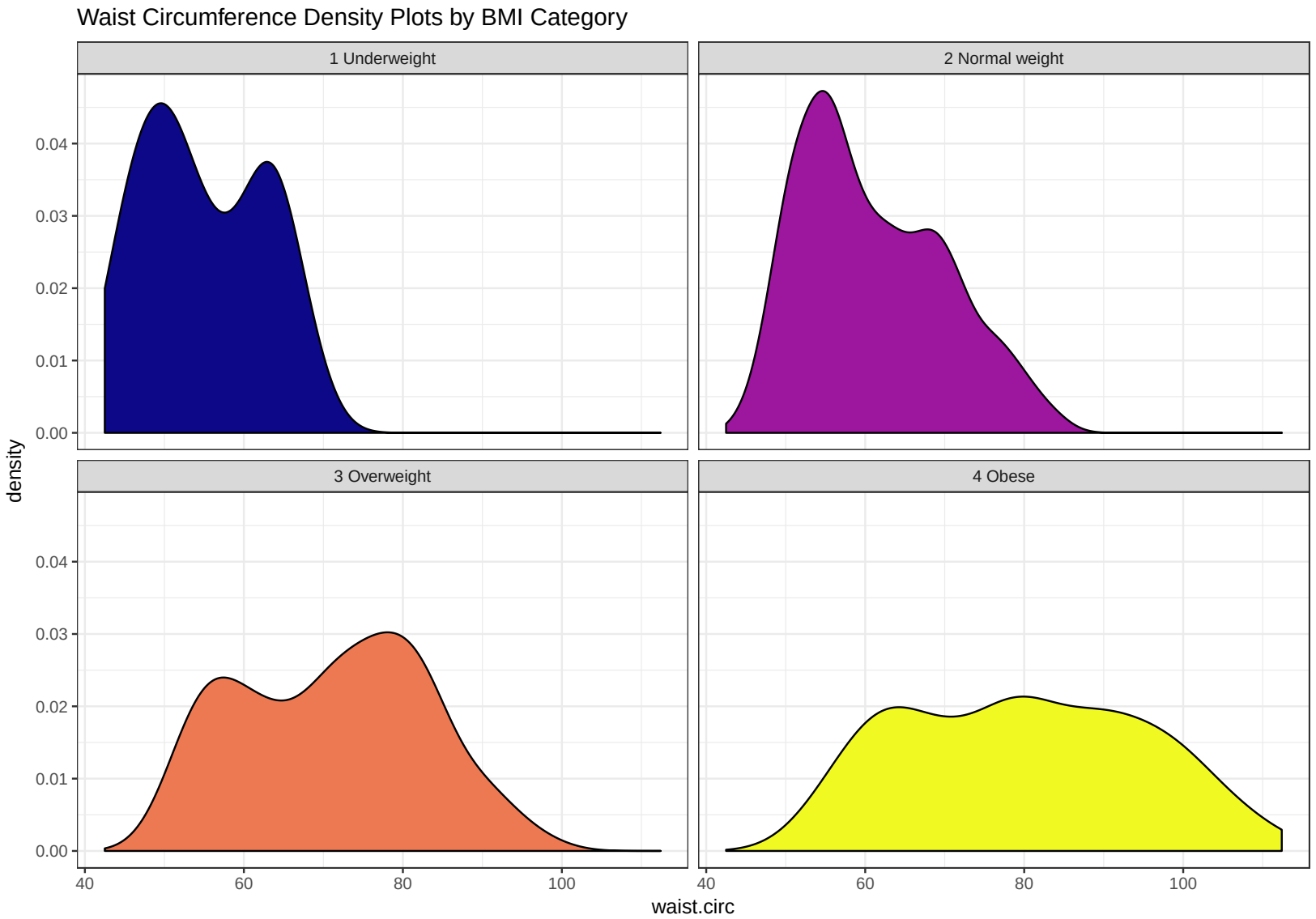
Comparison Boxplots with Notches



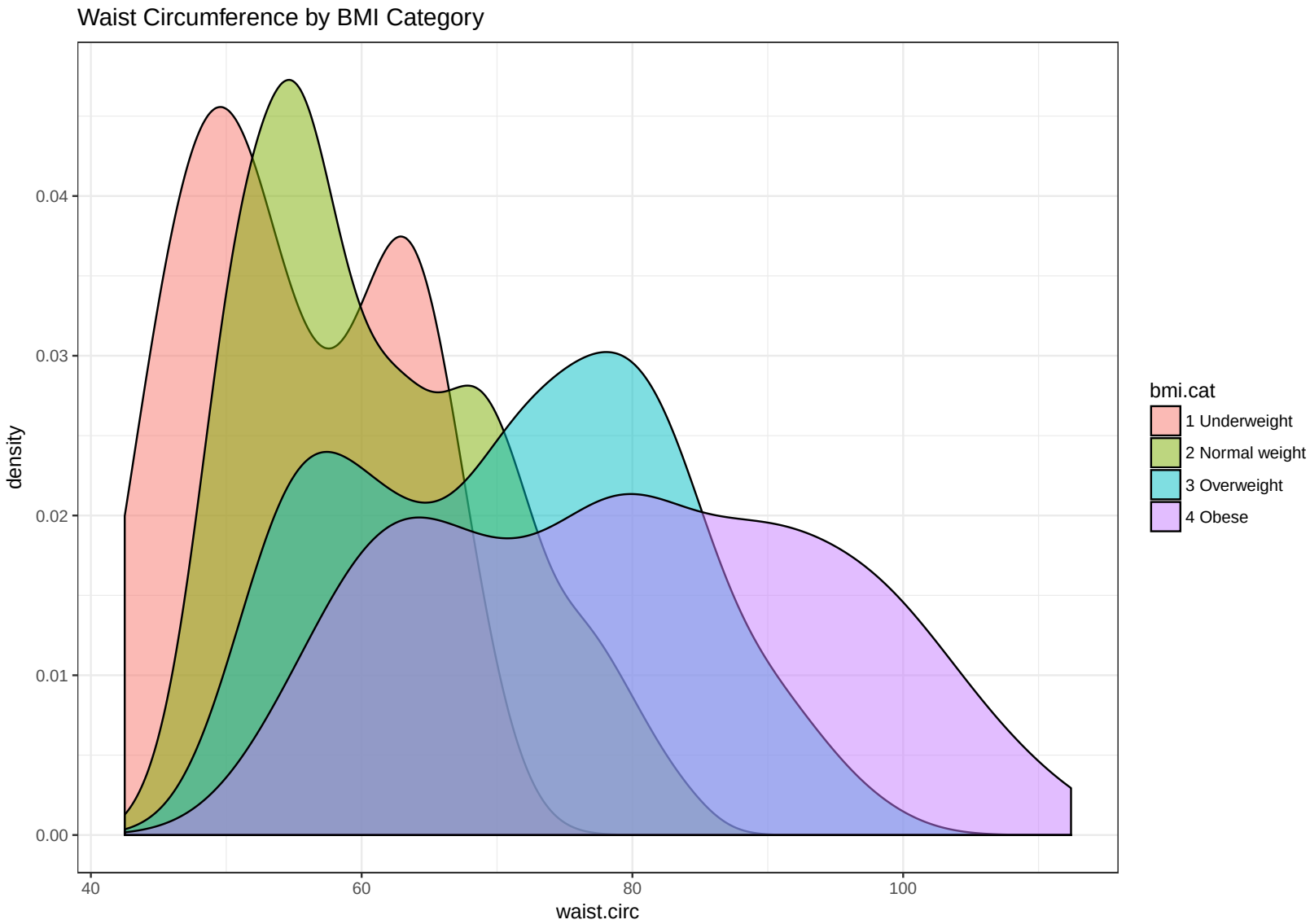
Comparing Distributions with Faceted Histograms



Density Plots, Faceted



Density Plots, Overlapping



Ridgeline Plots (formerly Joy Plots)