# HW 1

TJ Sipin

2022-06-25

# Problem 1 (Nile River data)

a.

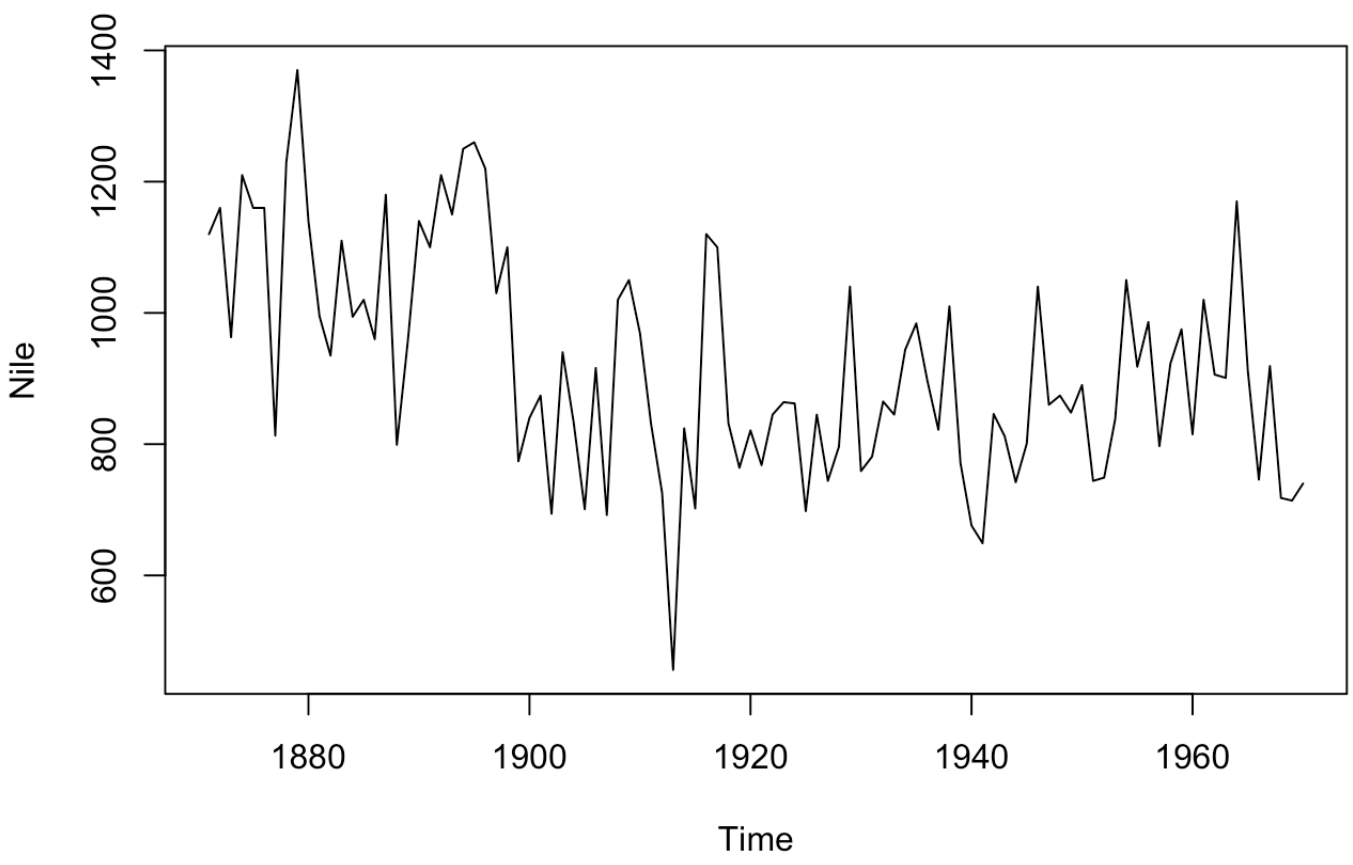
```
stem(Nile, scale = 1)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    4 | 6
##    5 |
##    6 | 5899
##    7 | 000123444455667778
##    8 | 000011222233344555556667779
##    9 | 0011222244466678899
##   10 | 0122234455
##   11 | 00012244566678
##   12 | 112356
##   13 | 7
```

The median flow quantity (in $10^8 m^3$) at Aswan in the Nile River is around the 800s and is a little right-skewed.

b.

```
plot(Nile)
```




The time series plot confirms that the median is around the 800s.

# Problem 2 (Curvature comparison)

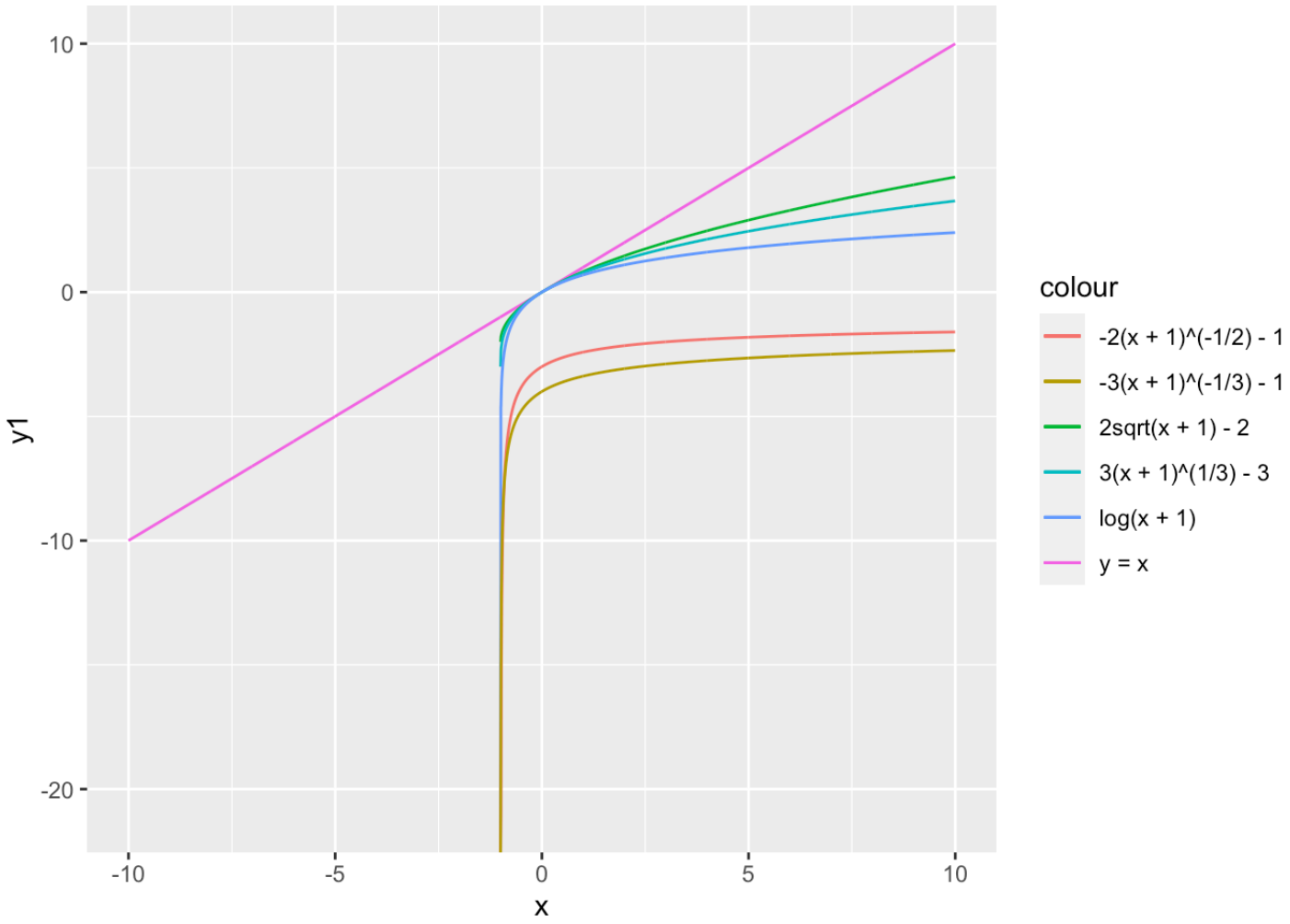
Draw curves of the following functions:

$$y = x, \quad y = 2\sqrt{x+1} - 2, \quad y = 3(x+1)^{1/3} - 3, \quad y = \log(x+1), \quad y = -2(x+1)^{-1/2} - 1, \quad y = -3(x+1)^{-1/3} - 1$$

```
## Warning in sqrt(x + 1): NaNs produced
```

```
## Warning in log(x + 1): NaNs produced
```

```
## Warning: Removed 900 row(s) containing missing values (geom_path).
## Removed 900 row(s) containing missing values (geom_path).
## Removed 900 row(s) containing missing values (geom_path).
## Removed 900 row(s) containing missing values (geom_path).
## Removed 900 row(s) containing missing values (geom_path).
```




The graph of the log function starts at x = -1 and increases strongly then levels out starting around x = 0.

# Problem 3 (Baby weight data)

```
baby <- read.delim("data/DISTRESS.DAT", header = FALSE) %>%
  mutate_at(.vars = 1:5,
            .funs = gsub,
            pattern = "\\*",
            replacement = "") %>%
  mutate_at(.vars = 1:5,
            .funs = as.numeric)
```

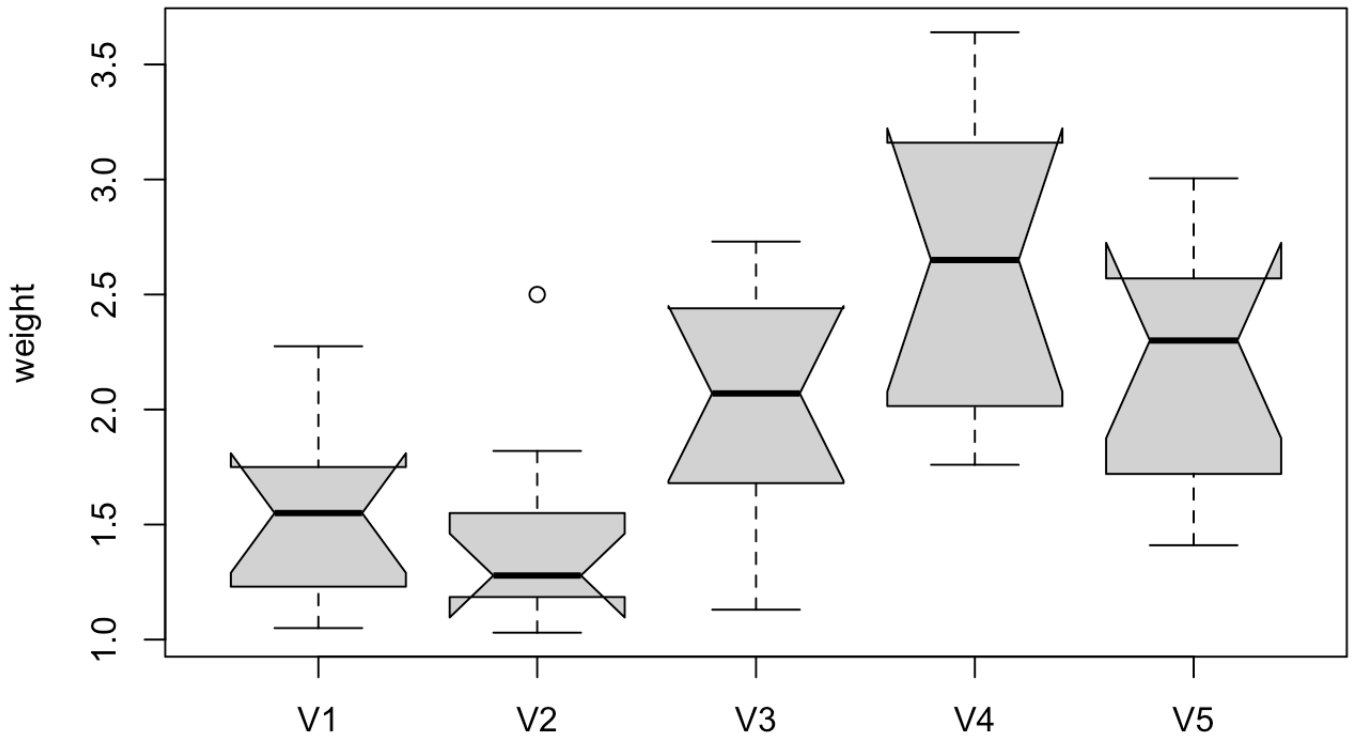
a.

```
baby_vector <- c(baby[[1]], baby[[2]], baby[[3]], baby[[4]], baby[[5]])
baby_stem <- stem(baby_vector, scale = 1)
```

```
##
##   The decimal point is at the |
##
##   1 | 0111222233334
##   1 | 566677778888999
##   2 | 001223344
##   2 | 56666778
##   3 | 0024
##   3 | 6
```

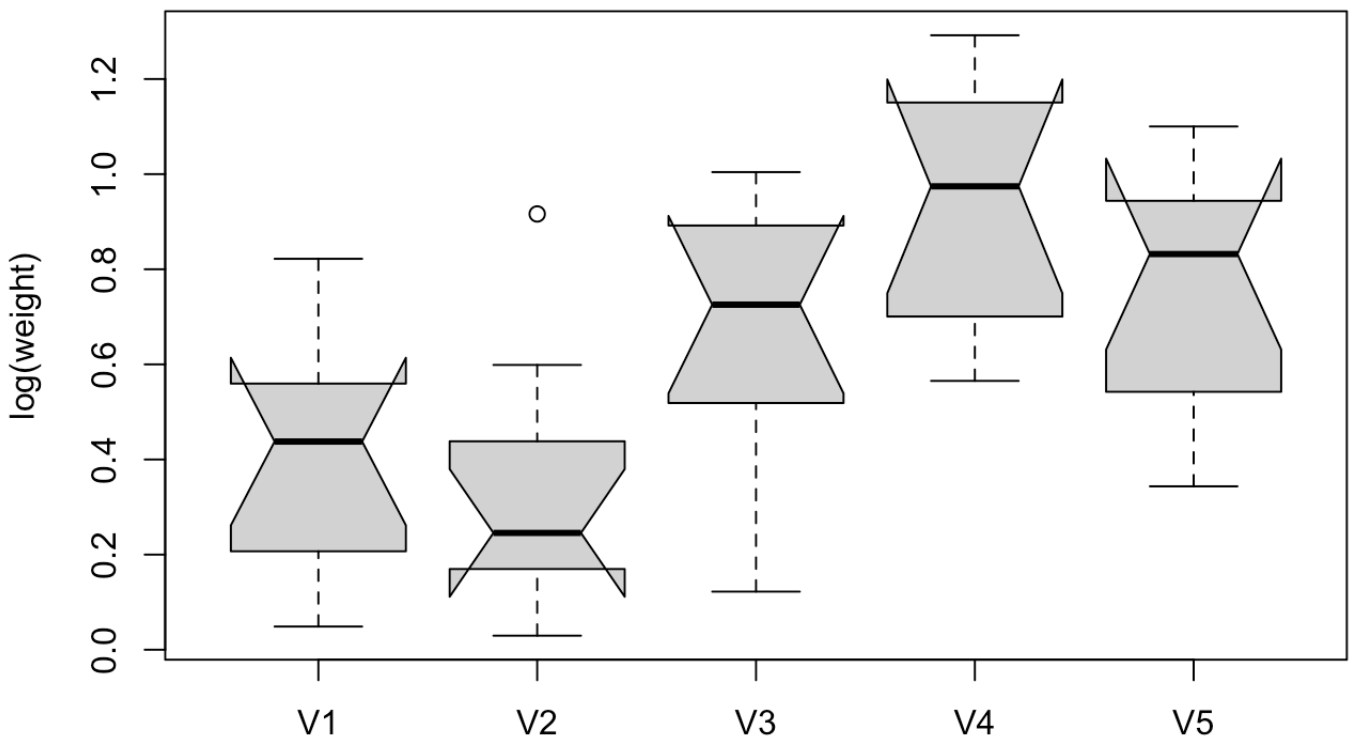
```
baby_box <- boxplot(baby, notch = T, ylab = "weight")
```

```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```




```
baby_t <- log(baby)
baby_t_box <- boxplot(baby_t, notch = T, ylab = "log(weight)")
```

```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```



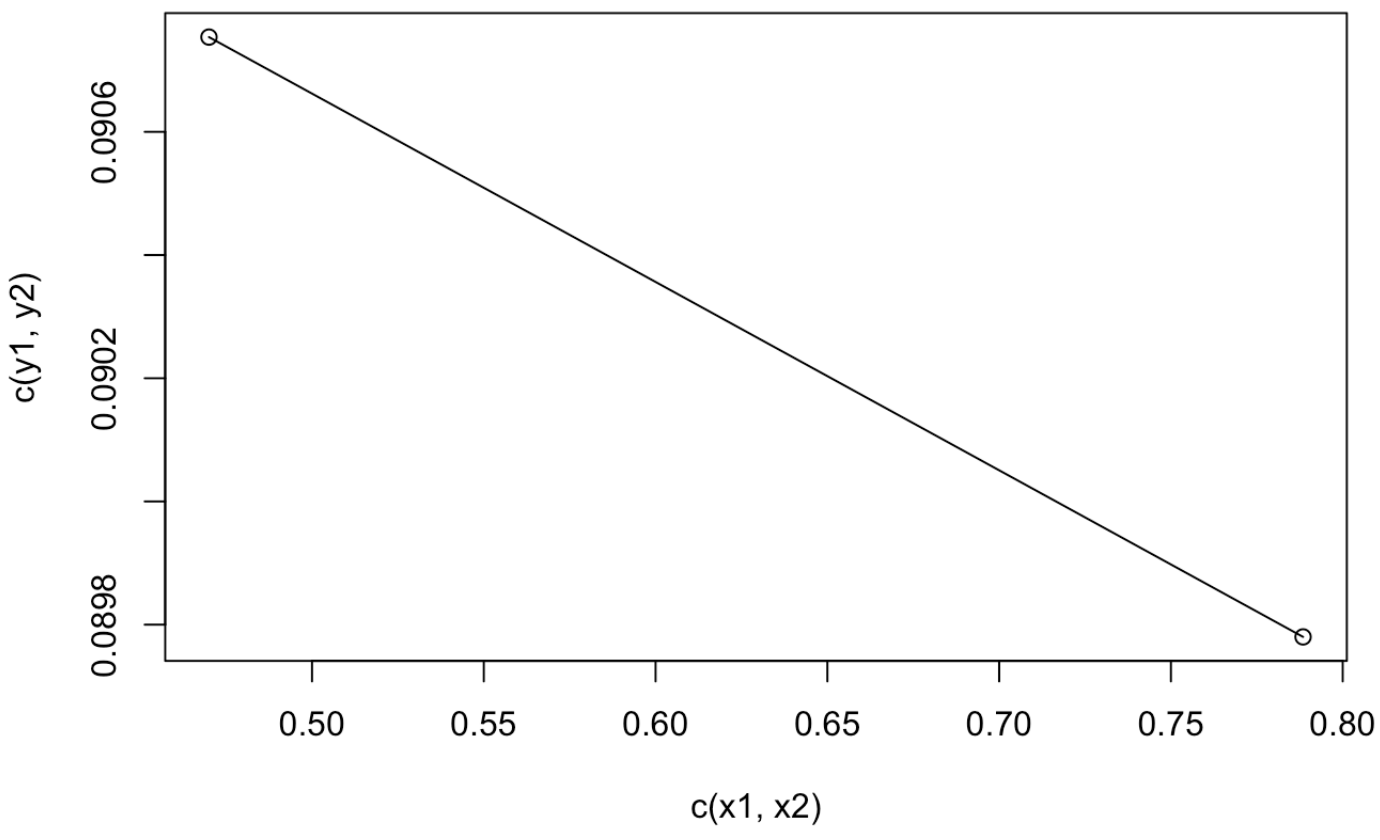

```
baby_t_vector <- c(baby_t[[1]], baby_t[[2]], baby_t[[3]], baby_t[[4]], baby_t[[5]])
baby_t_stem <- stem(baby_t_vector, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##    0 | 35
##    1 | 0267
##    2 | 013667
##    3 | 4
##    4 | 1457
##    5 | 2444677
##    6 | 0466
##    7 | 01499
##    8 | 2289
##    9 | 244469
##   10 | 048
##   11 | 05
##   12 | 29
```

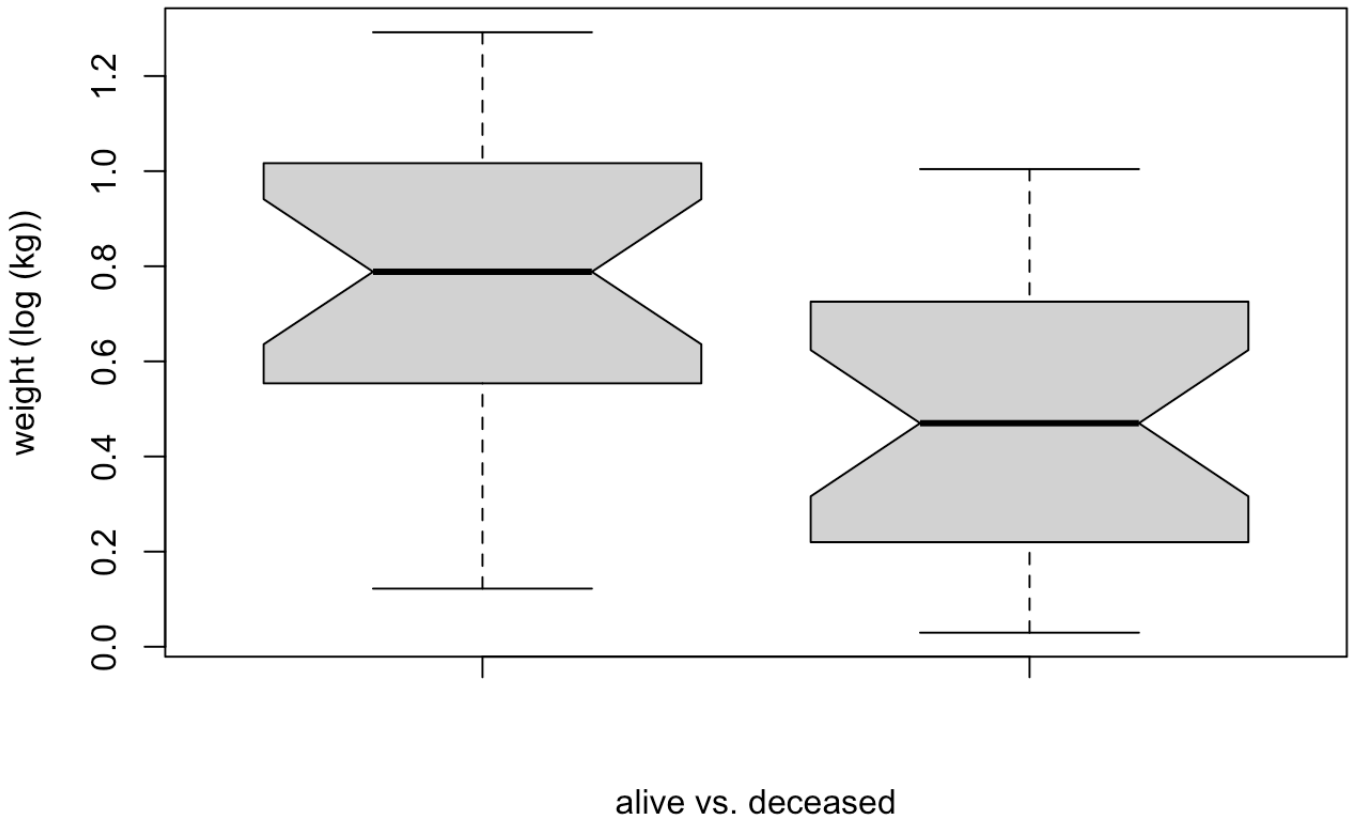
The boxplots do not look different, though the transformed data looks to have a wider distribution for each column. The stem-and-leaf displays are different however. Taking the log of the data causes the distribution to be less skewed.

b.

```
baby_raw <- read.delim("data/DISTRESS.DAT", header = FALSE)
baby_raw_vector <- c(baby_raw[[1]], baby_raw[[2]],
                     baby_raw[[3]], baby_raw[[4]],
                     baby_raw[[5]])
baby_deceased <- grep('\\*', baby_raw_vector, value = T)
baby_deceased <- baby_deceased %>%
  gsub(pattern = "\\*",
       replacement = "") %>%
  as.numeric()
baby_alive <- baby[baby != baby_deceased]

x1 <- median(log(baby_deceased))
y1 <- var(log(baby_deceased))

x2 <- median(log(baby_alive))
y2 <- var(log(baby_alive))

plot(c(x1, x2), c(y1, y2))
lines(c(x1, x2), c(y1, y2))
```



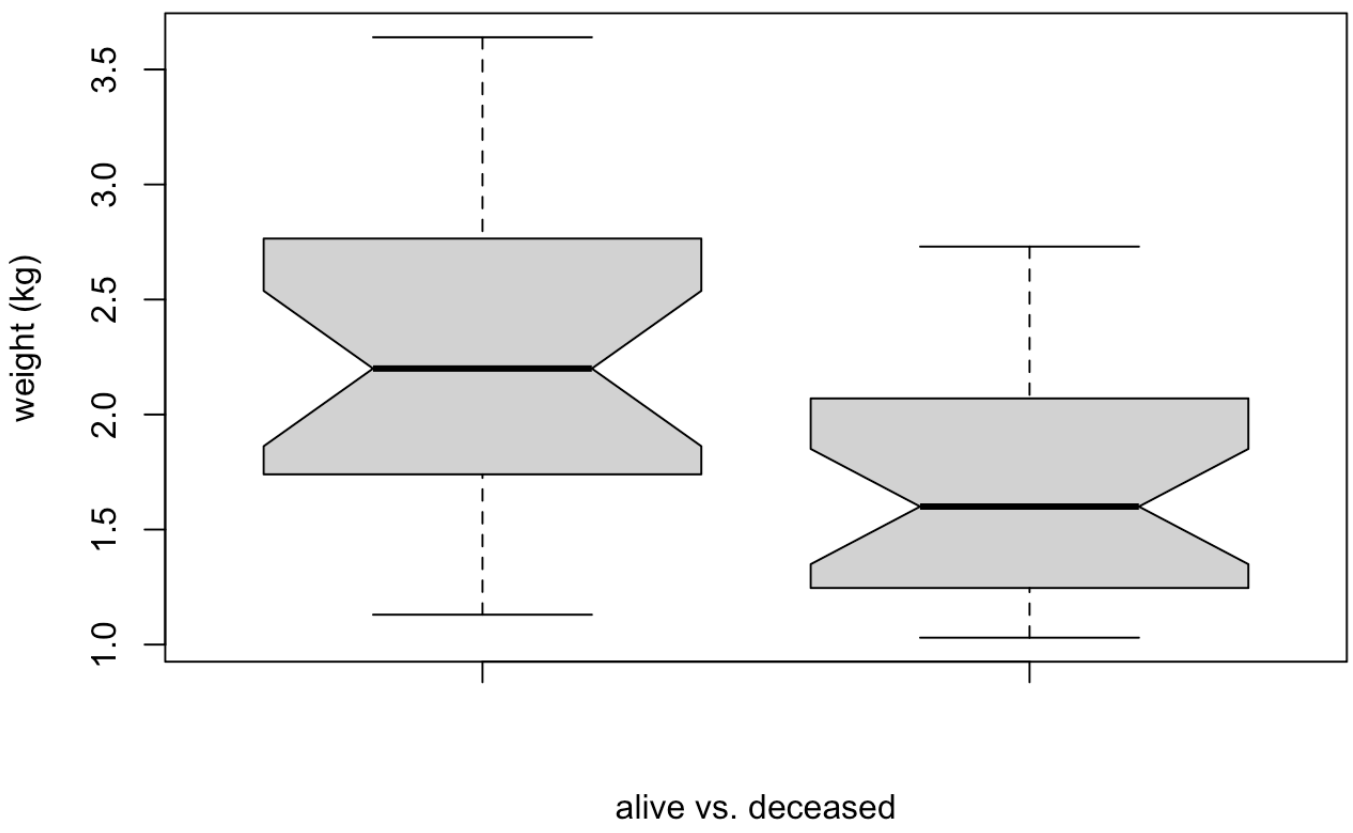

```
b = -0.003058
p = 0 # use log to transform data

boxplot(log(baby_alive), log(baby_deceased), notch = T,
        ylab = "weight (log (kg))",
        xlab = "alive vs. deceased")
```




After finding the value of $b$ to equal 0, $p = 0$ and we use log to transform the weight. The notches do not seem to overlap, so they are not significantly different (at a roughly 95% level).

c.

```
boxplot((baby_alive), (baby_deceased), notch = T,
        ylab = "weight (kg)",
        xlab = "alive vs. deceased")
```




```
stem(baby_alive)
```

```
##
##   The decimal point is at the |
##
##   1 | 14677789
##   2 | 0012466678
##   3 | 00246
```

```
stem(baby_deceased, scale = .4)
```

```
##
##   The decimal point is at the |
##
##   1 | 01122223333
##   1 | 566788899
##   2 | 2334
##   2 | 567
```

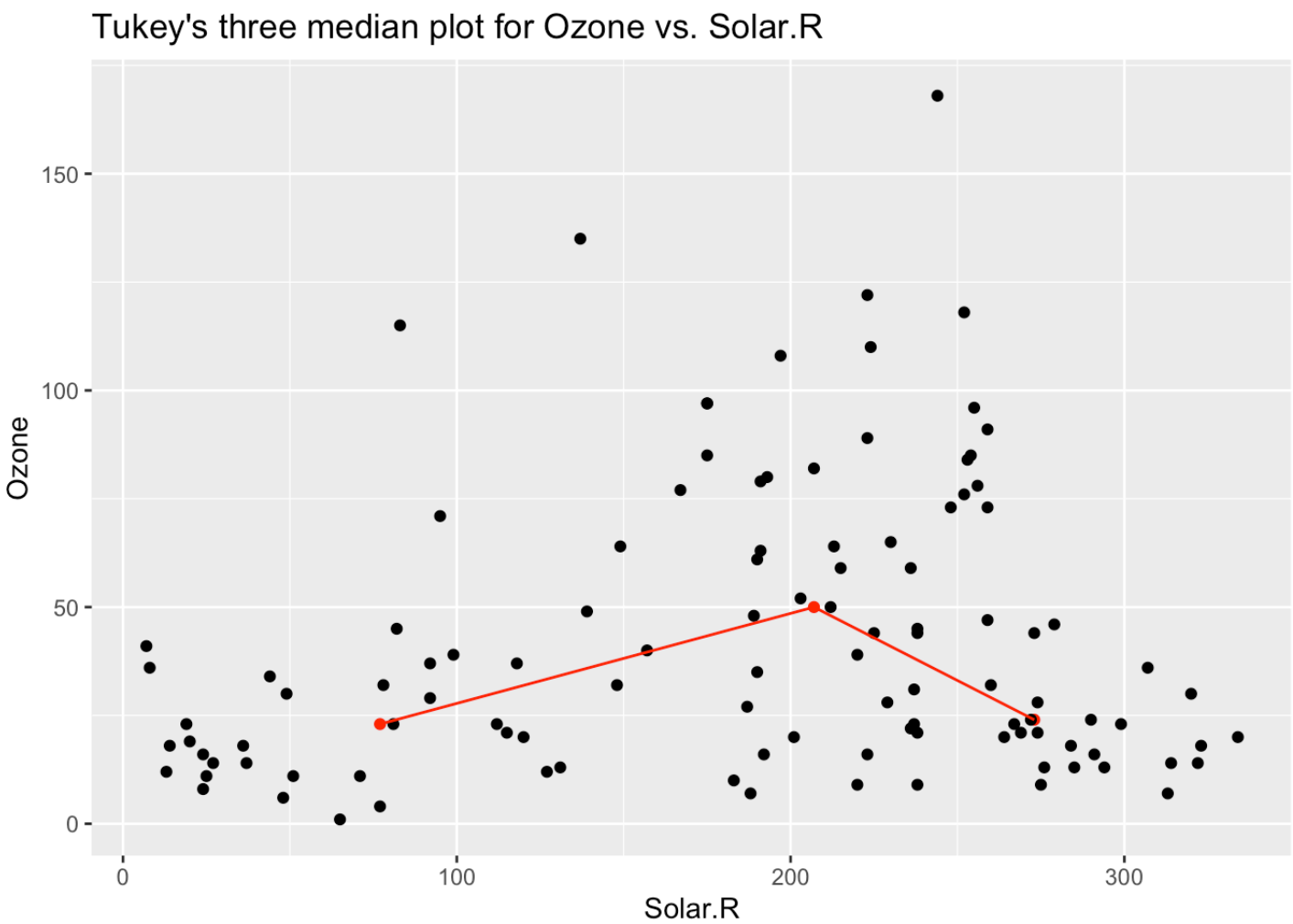
A specific weight a baby with severe idiopathic respiratory disease may die is around 1.5 to 1.6 kg, though may be even lower (around 1.4 to 1.8)

# Problem 4 (airquality data)

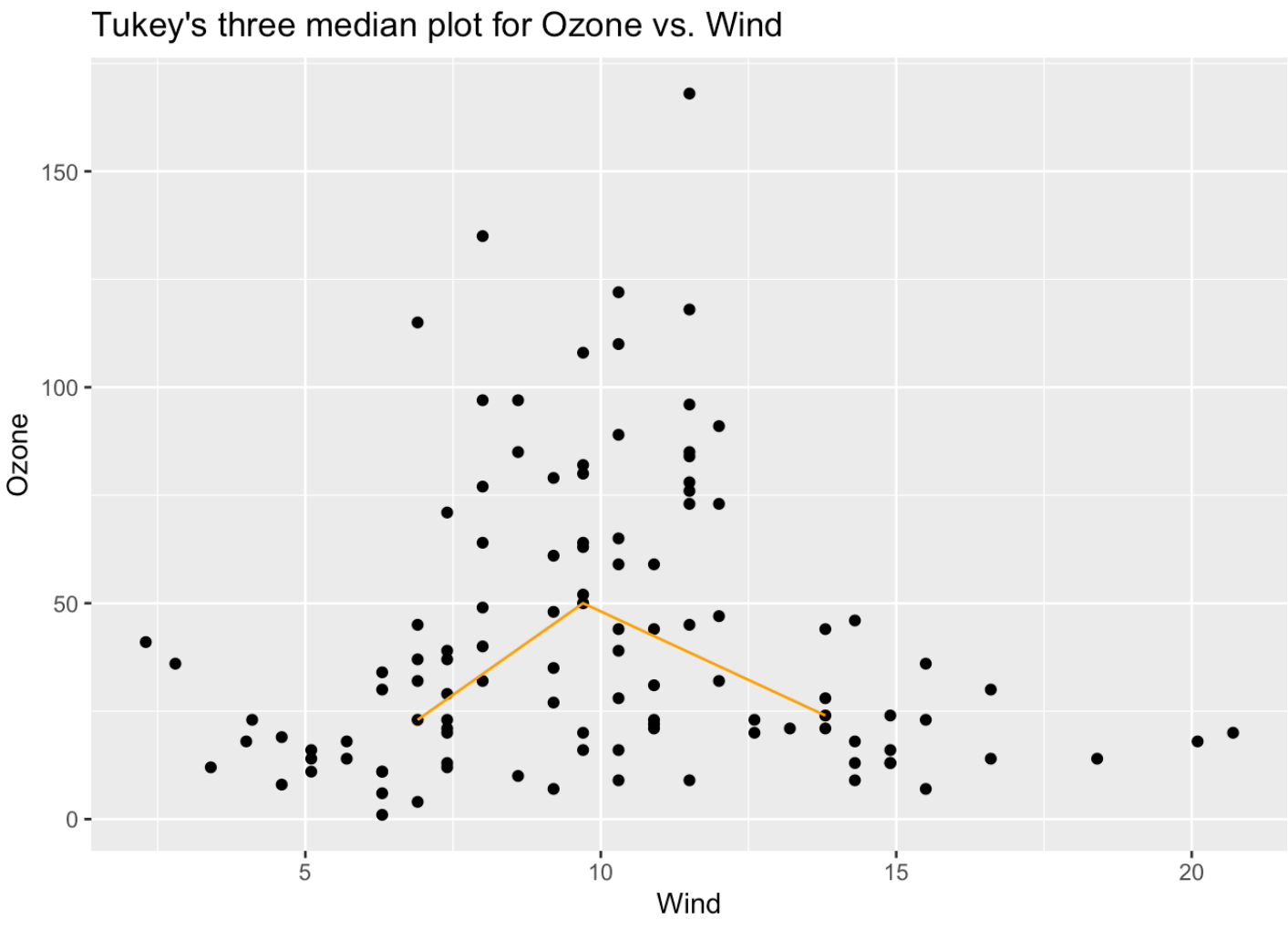
```
air <- airquality %>%
  select(c('Ozone', 'Solar.R', 'Wind', 'Temp')) %>%
  drop_na()
```

a.

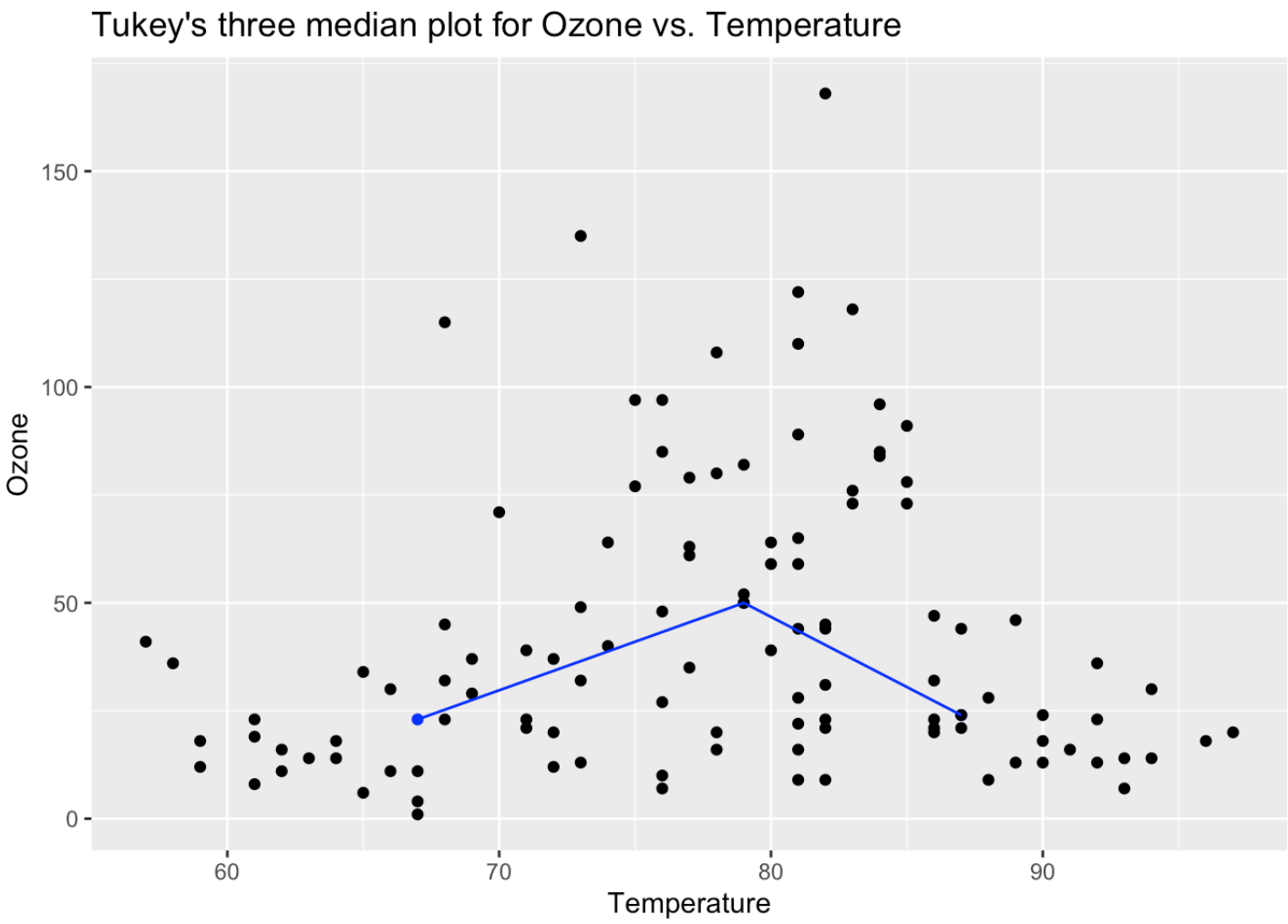
```
air_Solar.R <- air %>%
  select(c(Ozone, Solar.R))
air_Wind <- air %>%
  select(c(Ozone, Wind))
air_Temp <- air %>%
  select(c(Ozone, Temp))

air_Solar.R$Solar.R <-
  sort(air_Solar.R$Solar.R, decreasing = F)
air_Wind$Wind <-
  sort(air_Wind$Wind, decreasing = F)
air_Temp$Temp <-
  sort(air_Temp$Temp, decreasing = F)



y1L <- air_Solar.R$Ozone[1:37] %>% median()
y1M <- air_Solar.R$Ozone[38:74] %>% median()
y1R <- air_Solar.R$Ozone[75:111] %>% median()

y2L <- air_Wind$Ozone[1:37] %>% median()
y2M <- air_Wind$Ozone[38:74] %>% median()
y2R <- air_Wind$Ozone[75:111] %>% median()

y3L <- air_Temp$Ozone[1:37] %>% median()
y3M <- air_Temp$Ozone[38:74] %>% median()
y3R <- air_Temp$Ozone[75:111] %>% median()

x1L <- air_Solar.R$Solar.R[1:37] %>% median()
x1M <- air_Solar.R$Solar.R[38:74] %>% median()
x1R <- air_Solar.R$Solar.R[75:111] %>% median()

x2L <- air_Wind$Wind[1:37] %>% median()
x2M <- air_Wind$Wind[38:74] %>% median()
x2R <- air_Wind$Wind[75:111] %>% median()

x3L <- air_Temp$Temp[1:37] %>% median()
x3M <- air_Temp$Temp[38:74] %>% median()
x3R <- air_Temp$Temp[75:111] %>% median()

ggplot() +
  geom_point(aes(x = c(x1L, x1M, x1R),
                 y = c(y1L, y1M, y1R)),
             color = 'red') +
  geom_point(aes(x = air_Solar.R$Solar.R,
                 y = air_Solar.R$Ozone)) +
  geom_line(aes(x = c(x1L, x1M, x1R),
                 y = c(y1L, y1M, y1R)),
            color = 'red') +
  ggtitle('Tukey\'s three median plot for Ozone vs. Solar.R') +
  xlab('Solar.R') +
  ylab('Ozone')
```




```
ggplot() +
  geom_point(aes(x = c(x2L, x2M, x2R),
                 y = c(y2L, y2M, y2R)),
             color = 'orange') +
  geom_point(aes(x = air_Wind$Wind,
                 y = air_Wind$Ozone)) +
  geom_line(aes(x = c(x2L, x2M, x2R),
                 y = c(y2L, y2M, y2R)),
             color = 'orange') +
  ggtitle('Tukey\'s three median plot for Ozone vs. Wind') +
  xlab('Wind') +
  ylab('Ozone')
```

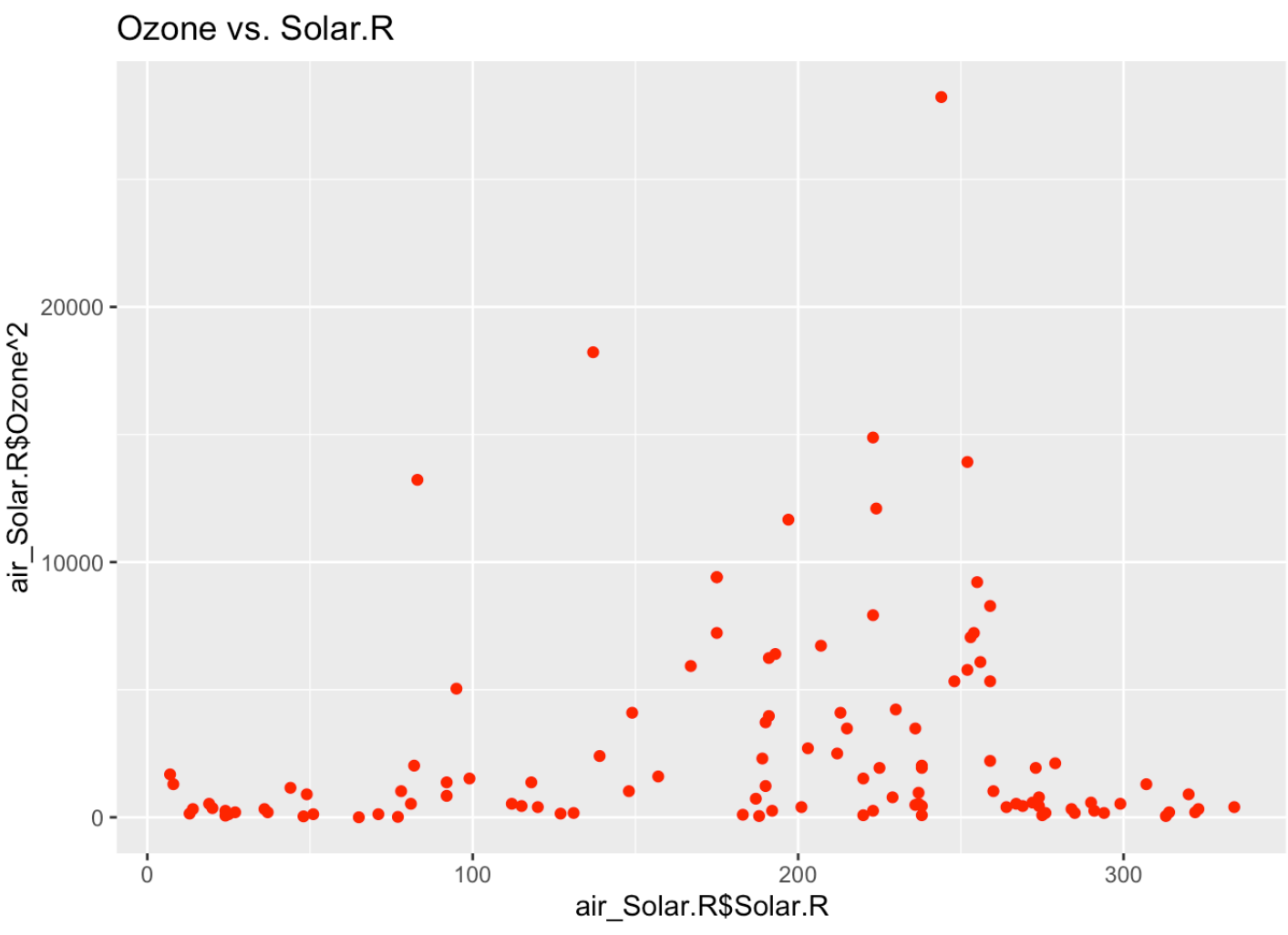


```
ggplot() +
  geom_point(aes(x = c(x3L, x3M, x3R),
                 y = c(y3L, y3M, y3R)),
             color = 'blue') +
  geom_point(aes(x = air_Temp$Temp,
                 y = air_Temp$Ozone)) +
  geom_line(aes(x = c(x3L, x3M, x3R),
                 y = c(y3L, y3M, y3R)),
             color = 'blue') +
  ggtitle('Tukey\'s three median plot for Ozone vs. Temperature') +
  xlab('Temperature') +
  ylab('Ozone')
```

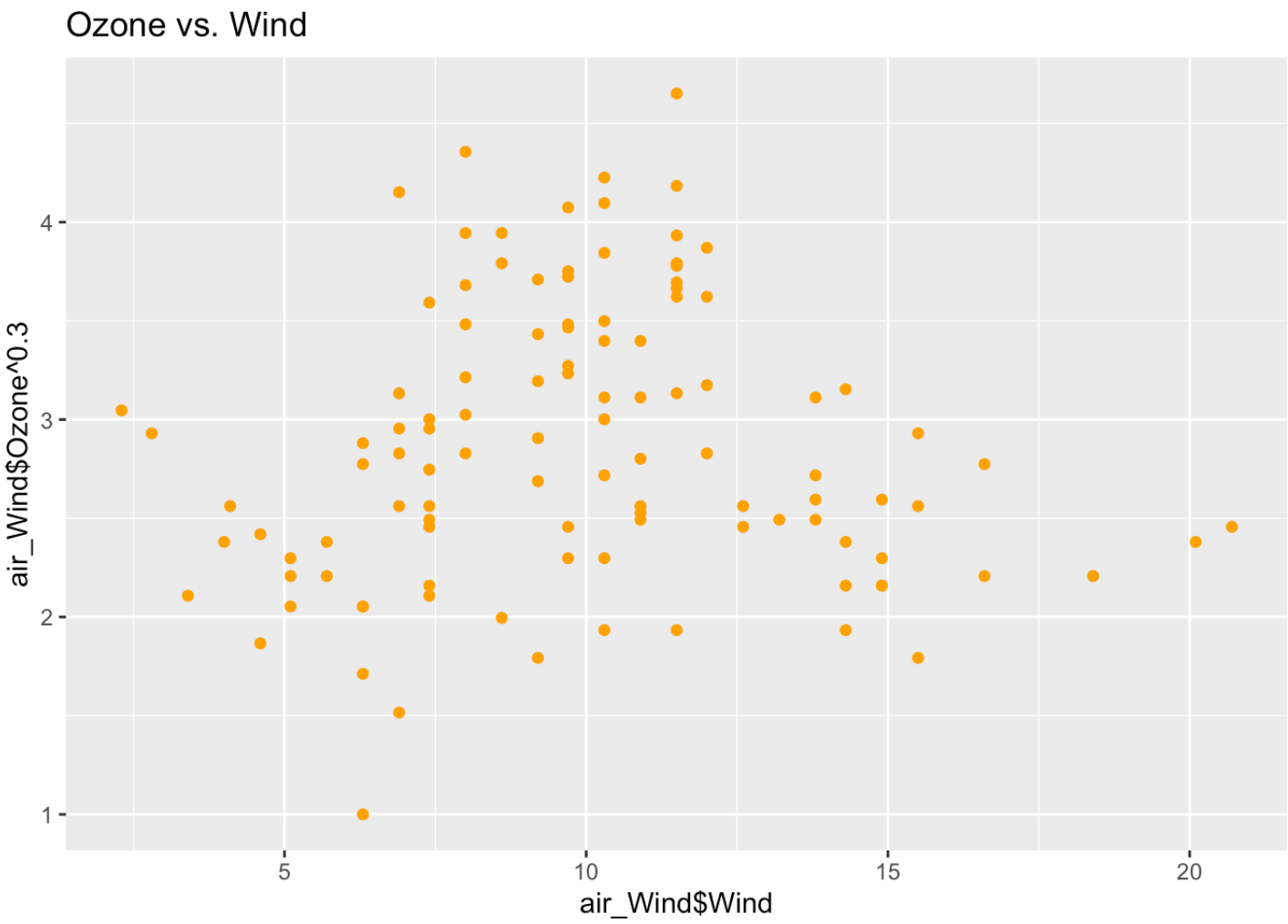



b.

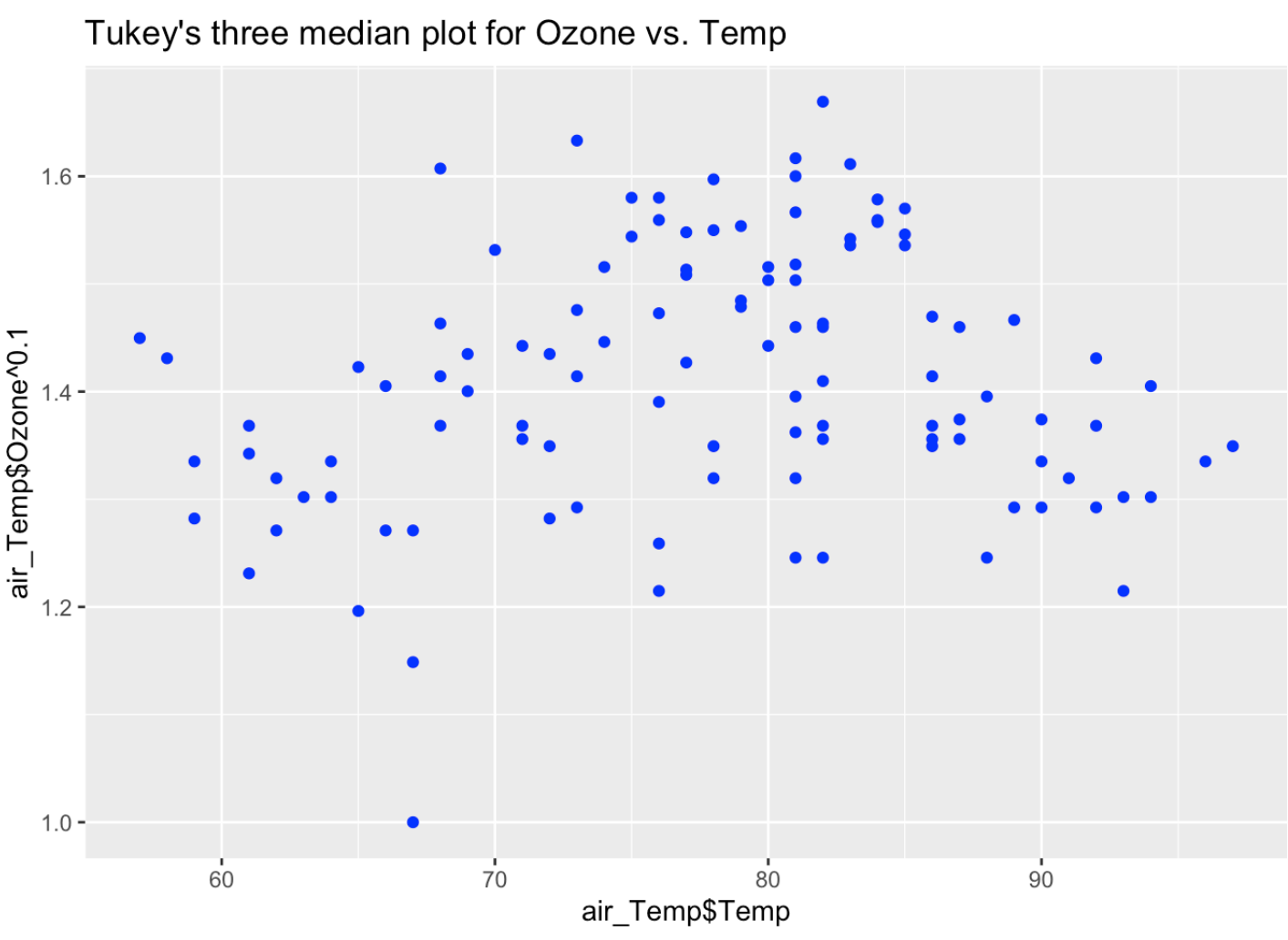
```
ggplot() +
  geom_point(aes(x = air_Solar.R$Solar.R,
                 y = air_Solar.R$Ozone^2),
             color = 'red') +
  ggtitle('Ozone vs. Solar.R')
```



```
ggplot() +
  geom_point(aes(x = air_Wind$Wind,
                 y = air_Wind$Ozone^0.3),
             color = 'orange') +
  ggtitle('Ozone vs. Wind')
```



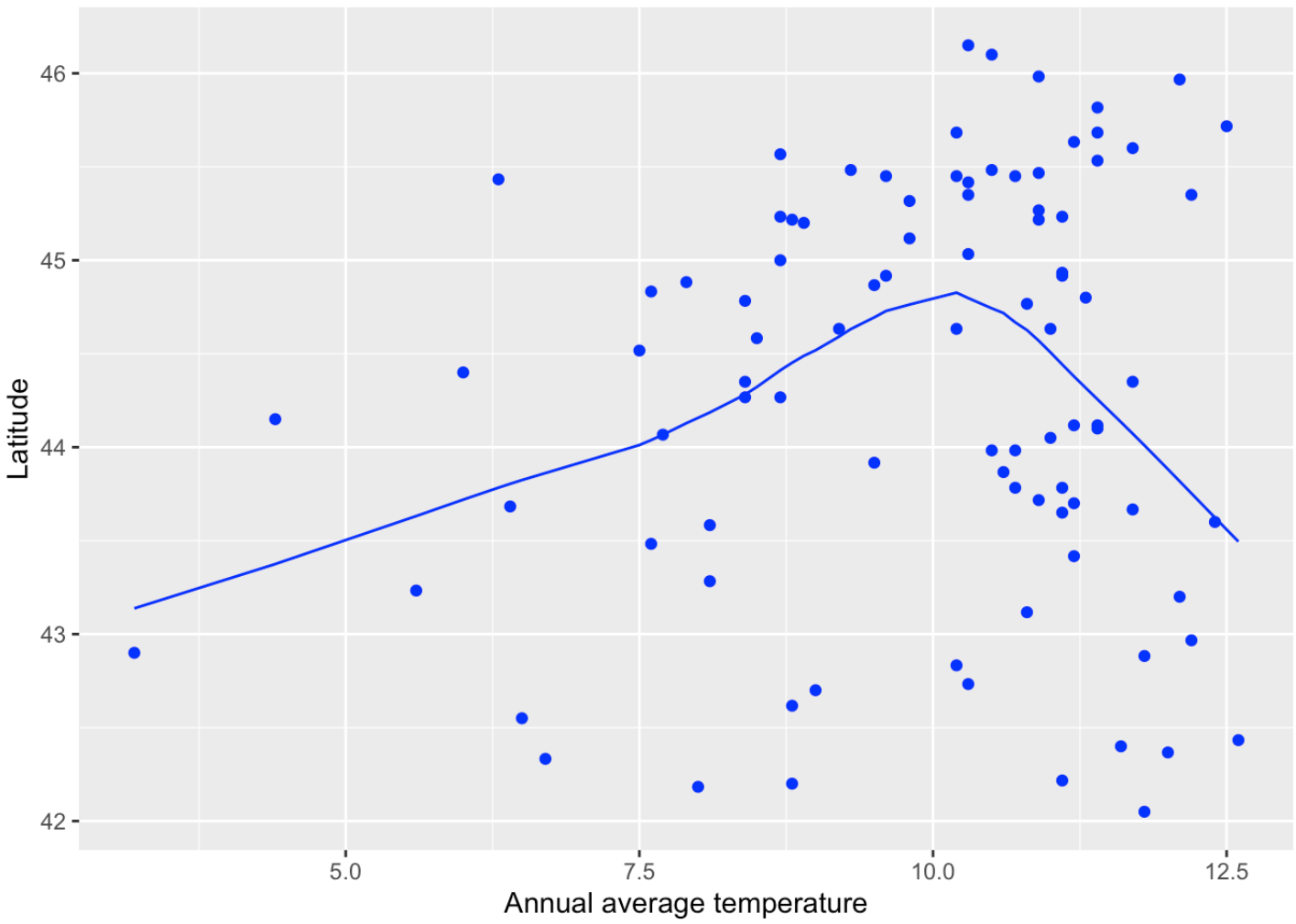

```
ggplot() +
  geom_point(aes(x = air_Temp$Temp,
                 y = air_Temp$Ozone^0.1),
             color = 'blue') +
  ggtitle('Tukey\'s three median plot for Ozone vs. Temp')
```



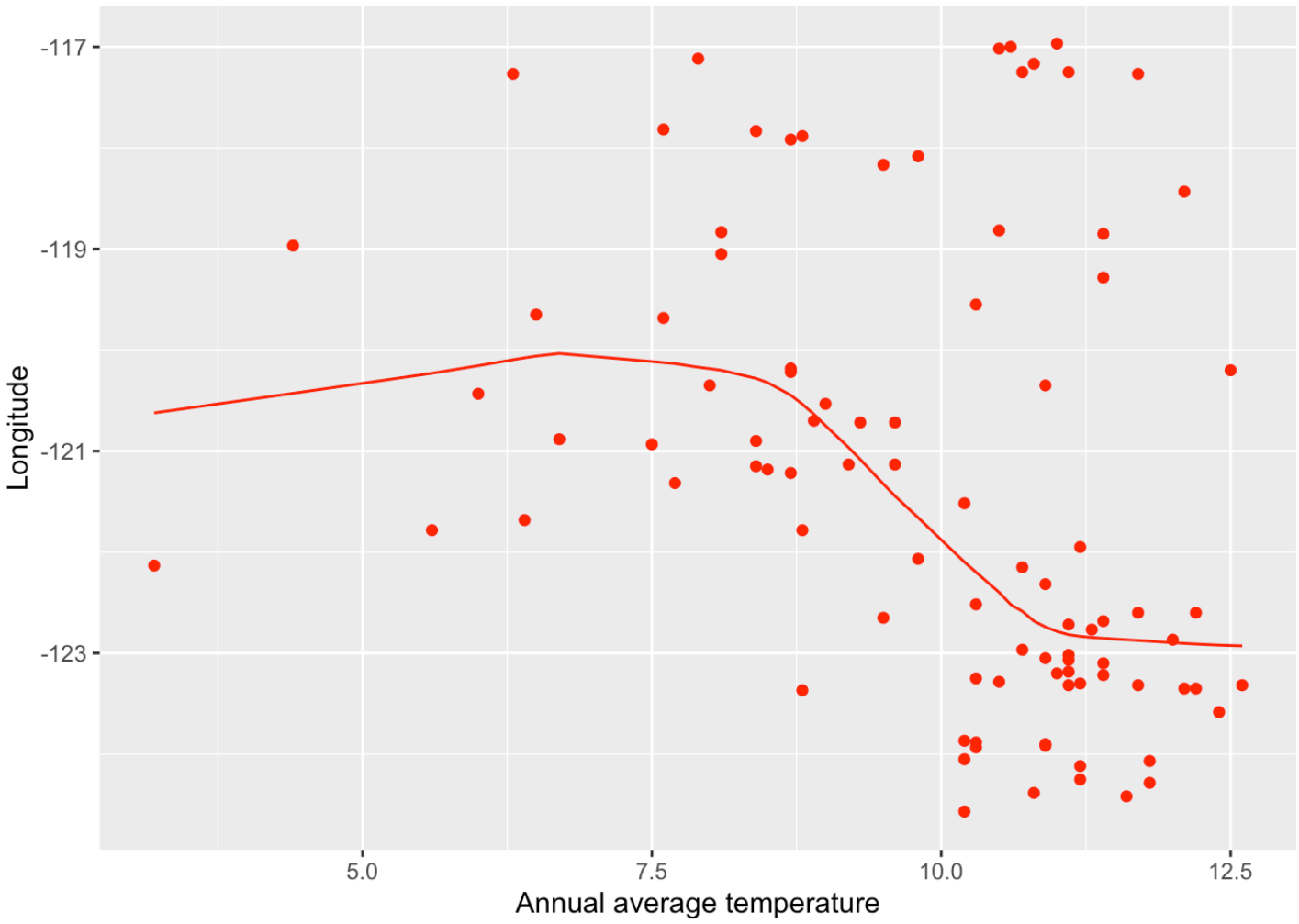

The notes said to lower the power of $x$, but I feel like it's better to lower the power of $y$. I am more satisfied with the 2nd and 3rd graphs (lower power of $y$) than with the 1st graph. They also look more homoscedastic.

# Problem 5 (Oregon temperature data).

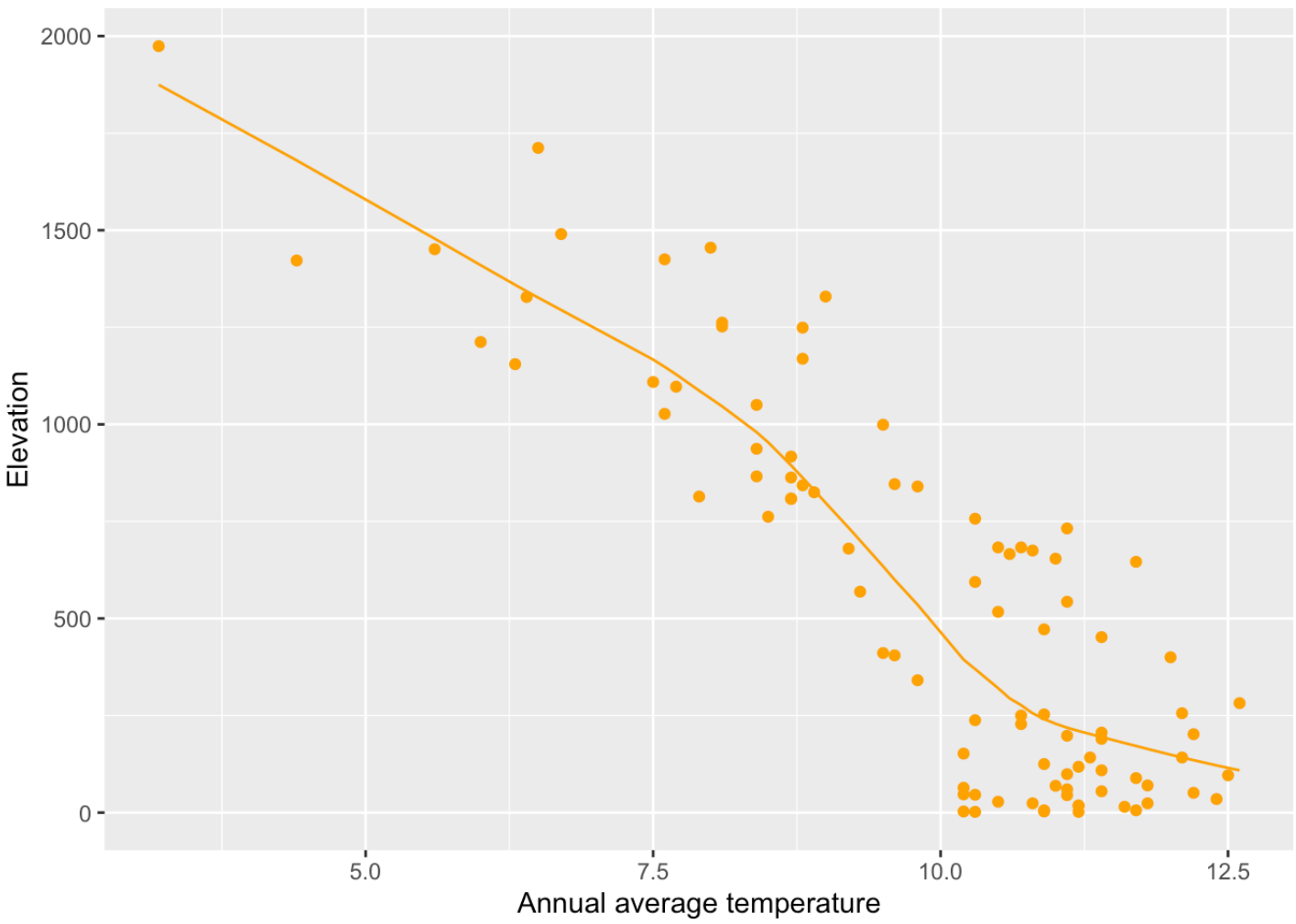
```
oregon <- read.csv('data/ortann.csv')

# tann vs. latitude
tann_lat <- lowess(x = oregon$tann,
                   y = oregon$latitude,
                   f = 2/3)
# tann vs. longitude
tann_long <- lowess(x = oregon$tann,
                    y = oregon$longitude,
                    f = 2/3)

# tann vs. elevation
tann_elev <- lowess(x = oregon$tann,
                    y = oregon$elevation,
                    f = 2/3)


ggplot() +
  geom_line(aes(x = tann_lat$x,
                y = tann_lat$y),
            color = 'blue') +
  geom_point(aes(x = oregon$tann,
                 y = oregon$latitude),
             color = 'blue') +
  xlab('Annual average temperature') +
  ylab('Latitude')
```




```
ggplot() +
  geom_line(aes(x = tann_long$x,
                y = tann_long$y),
            color = 'red') +
  geom_point(aes(x = oregon$tann,
                 y = oregon$longitude),
             color = 'red') +
  xlab('Annual average temperature') +
  ylab('Longitude')
```




```
ggplot() +
  geom_line(aes(x = tann_elev$x,
                y = tann_elev$y),
            color = 'orange') +
  geom_point(aes(x = oregon$tann,
                 y = oregon$elevation),
             color = 'orange') +
  xlab('Annual average temperature') +
  ylab('Elevation')
```




In Oregon, temperature tends to be pretty independent of latitude, though it tends to be slightly dependent of longitude as it increases going west. Elevation was most correlated with temperature, as temperature increases with a decrease in elevation.