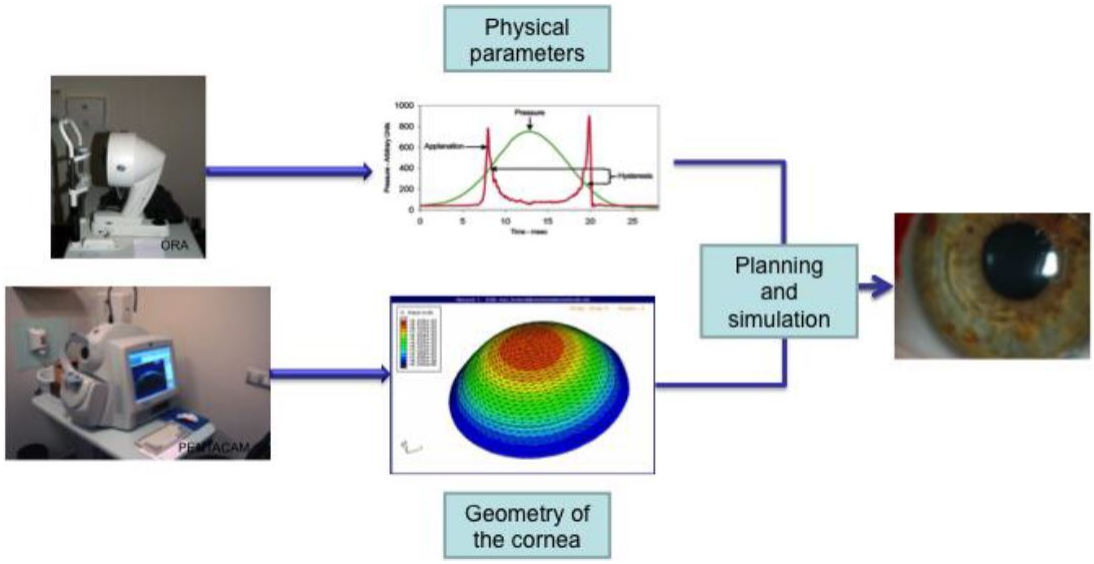
Practical 4: Visualisation using qplot()

Laura Rodriguez Navas

April 2020

Keratoconus is a disorder that affects the cornea through an abnormal growth of collagen fibres. This makes the cornea become conical with an important vision loss. There are many possible treatments, but one common solution is the insertion of intrastromal corneal ring segments, such that the cornea is flattened.



The dataset "queratocono.csv" includes information about 394 patients with Keratoconus who were treated with ring placement. The variables that were recorded are:

1. K1: keratometry or main corneal curvature.
2. K2: perpendicular curvature to K1.
3. Ch: corneal hysteresis.
4. Na: number of rings (1 or 2).
5. Incision: angle in which the cornea is cut.
6. Prof: depth of the incision.
7. Diam: diameter of the incision.
8. Grosor: Incision thickness.
9. Longitud1: Angle of placement of the first ring (surgical parameter).
10. Longitud2: Angle of placement of the second ring (surgical parameter).
11. grosor1: Thickness of the first ring.
12. grosor2: Thickness of the second ring.
13. long1: arc length of the first ring.
14. long2: arc length of the second ring.
15. K1.salida: keratometry or main corneal curvature after the placement of the ring(s).
16. Astig: astigmatism curvature after the placement of the ring(s) (K1.salida – K2.salida).

Check that there is no NA value in the dataset.

```
any(is.na(queratocono))
```

```
## [1] FALSE
```

The data in the dataset is sorted by column na and all the variables are numerical.

```
queratocono <- queratocono[order(queratocono$na), ]
str(queratocono)
```

```
## 'data.frame':    394 obs. of  16 variables:
##  $ K1       : num  45.7 44.2 44.2 53.1 40.7 35.4 51.5 40.6 49.5 44.9 ...
##  $ K2       : num  50.1 44.8 45.8 53.8 44.5 42 55.4 45.4 53.7 47.5 ...
##  $ ch       : num  11.1 8.9 11.1 7.5 9.4 8.2 10.1 8.6 7.2 7.6 ...
##  $ na       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Incision : int  30 180 130 60 30 170 165 30 180 60 ...
##  $ Prof     : int  370 327 380 400 477 420 400 400 434 367 ...
##  $ diam     : int  5 6 6 5 5 5 6 6 5 5 ...
##  $ grosor   : int  250 200 200 200 200 250 200 250 200 200 ...
##  $ Longitud1: int  160 150 210 160 160 160 150 150 210 160 ...
##  $ Longitud2: int  160 150 210 160 160 160 150 150 210 160 ...
##  $ grosor1  : int  250 200 200 200 200 250 200 250 200 200 ...
##  $ grosor2  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ long1    : int  160 150 210 160 160 160 150 150 210 160 ...
##  $ long2    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ K1.salida: num  45.8 42.8 43.1 48.7 39.8 35.6 48.7 42.7 47.4 43.7 ...
##  $ Astig    : num  1.9 2.3 3.9 2.8 3.8 4 5.3 1.2 0.3 1.8 ...
```

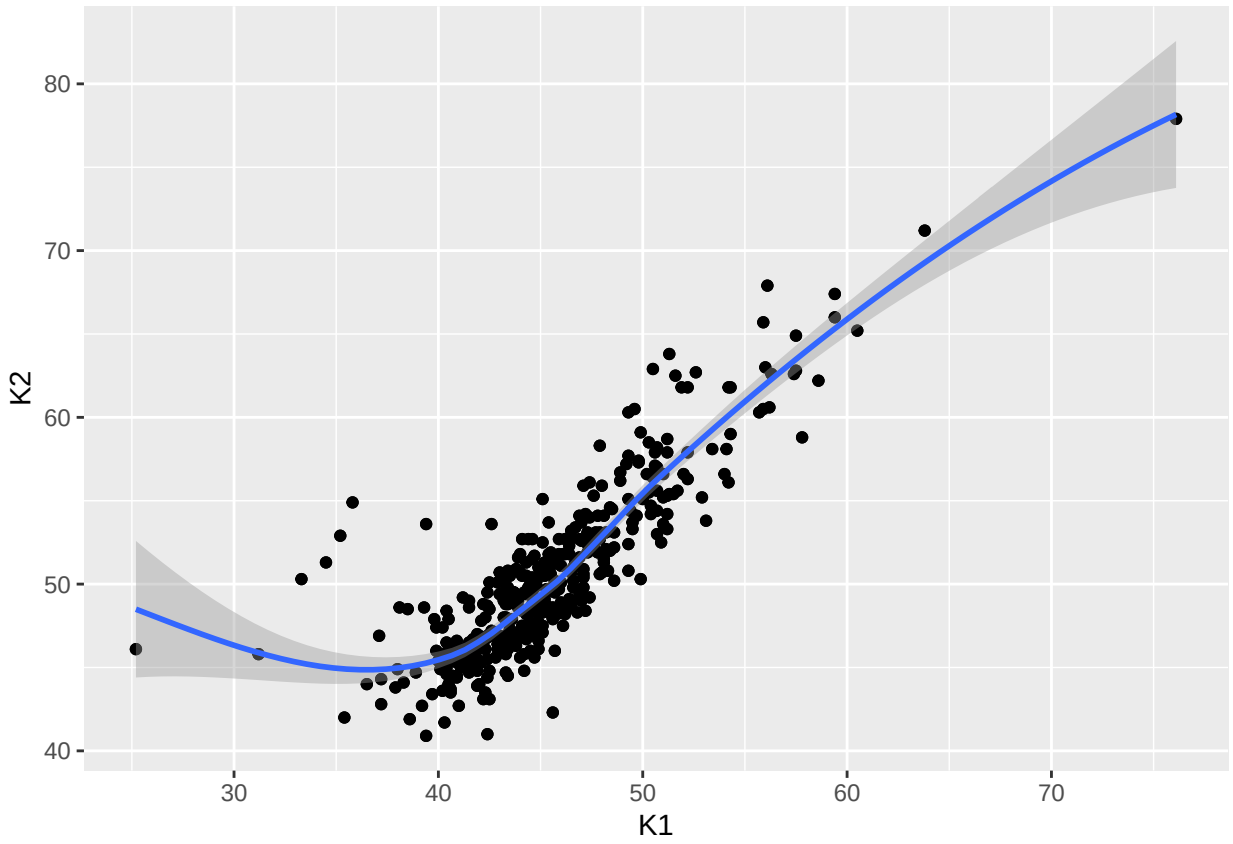
Data Visualisation

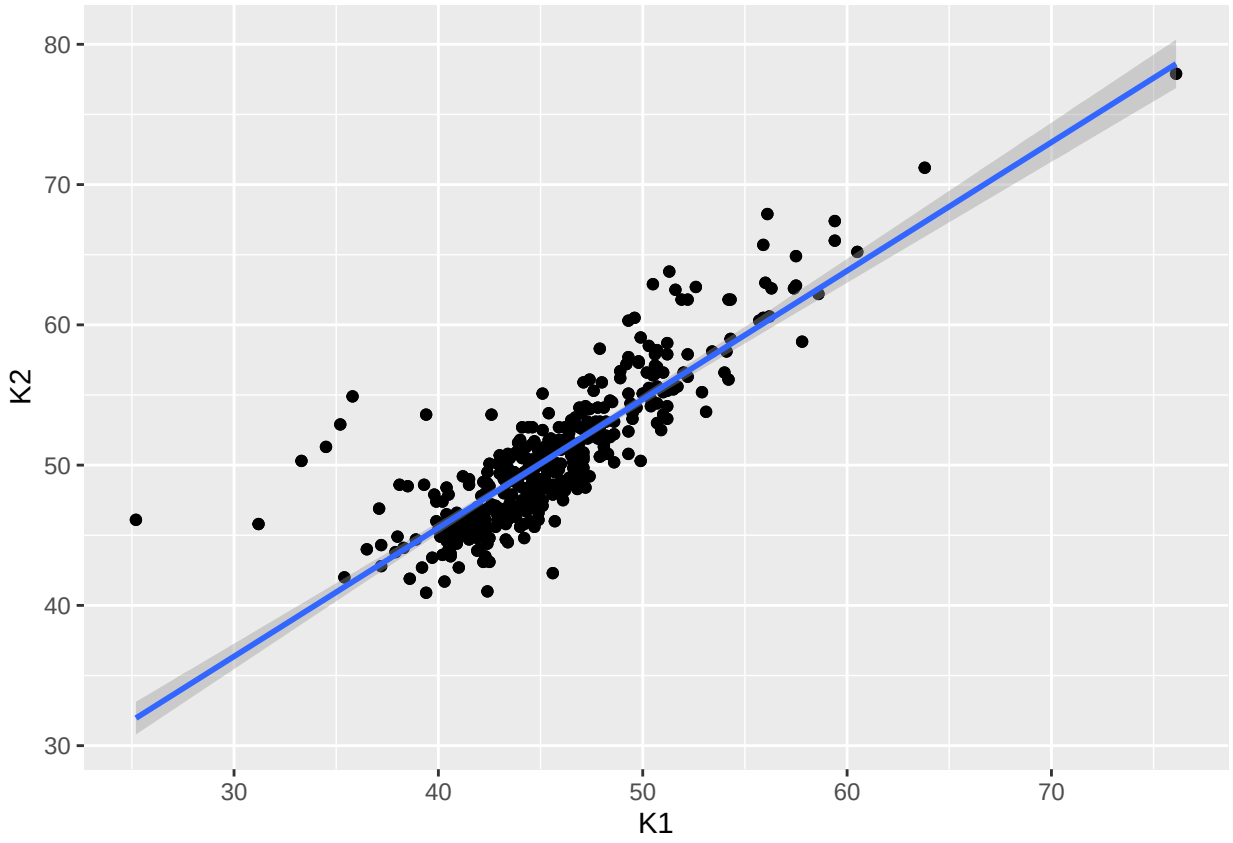
In order to analyse the information in a visual way:

1. Study the relation between K1 and K2 with smoother (by default and using linear regression).

```
qplot(K1, K2, data = queratocono) +
  geom_point() +
  geom_smooth(method = "loess") +
  xlab("K1") + ylab("K2")
```

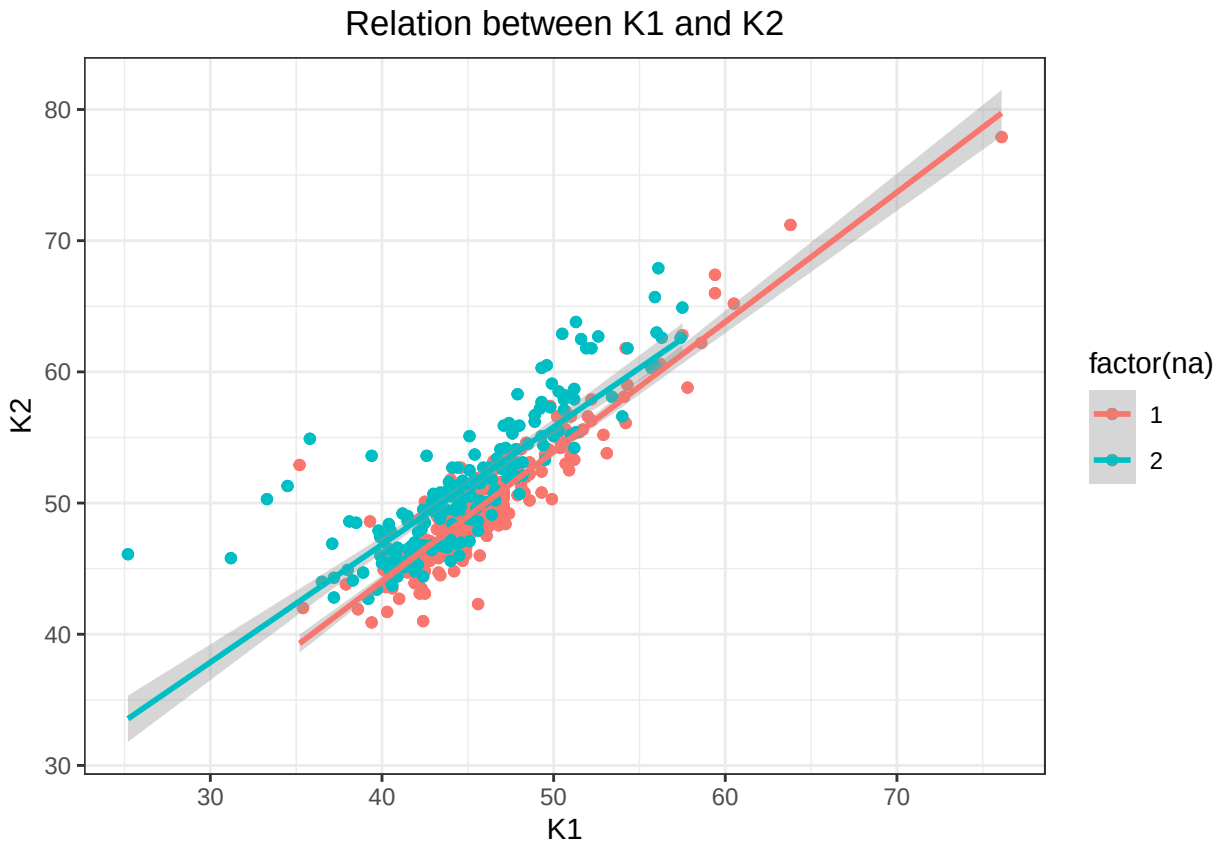
```
qplot(K1, K2, data = queratocono) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("K1") + ylab("K2")
```


80
70
60
50
40
K2
30
40
50
60
70
K1


80
70
60
50
40
30
K2
30
40
50
60
70
K1

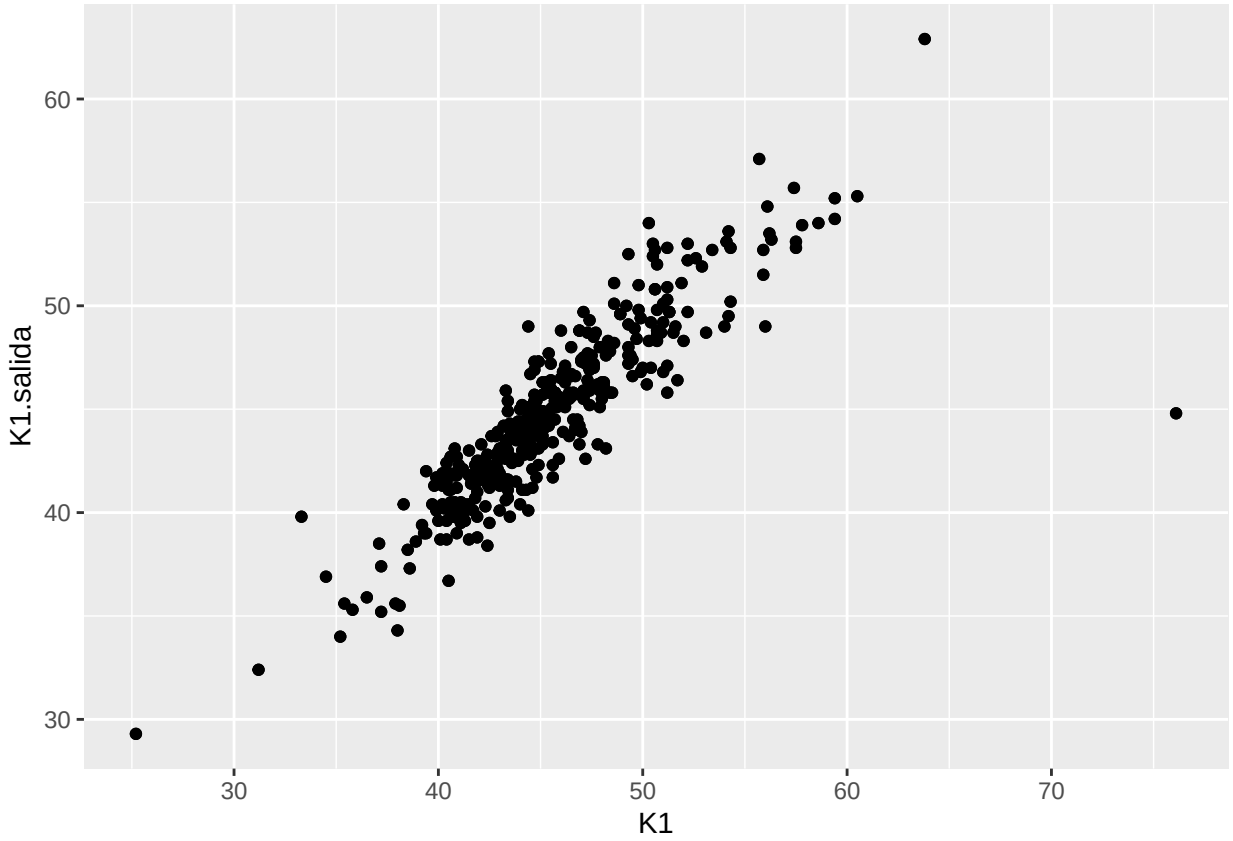
2. Study the relation between K1 and K2 distinguishing by factor na.

```
qplot(K1, K2, data = queratocono, colour = factor(na)) +
  geom_point() +
  geom_smooth(method = lm) +
  xlab("K1") + ylab("K2") +
  ggtitle("Relation between K1 and K2") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



3. Study the relation between K1 and K1.salida.

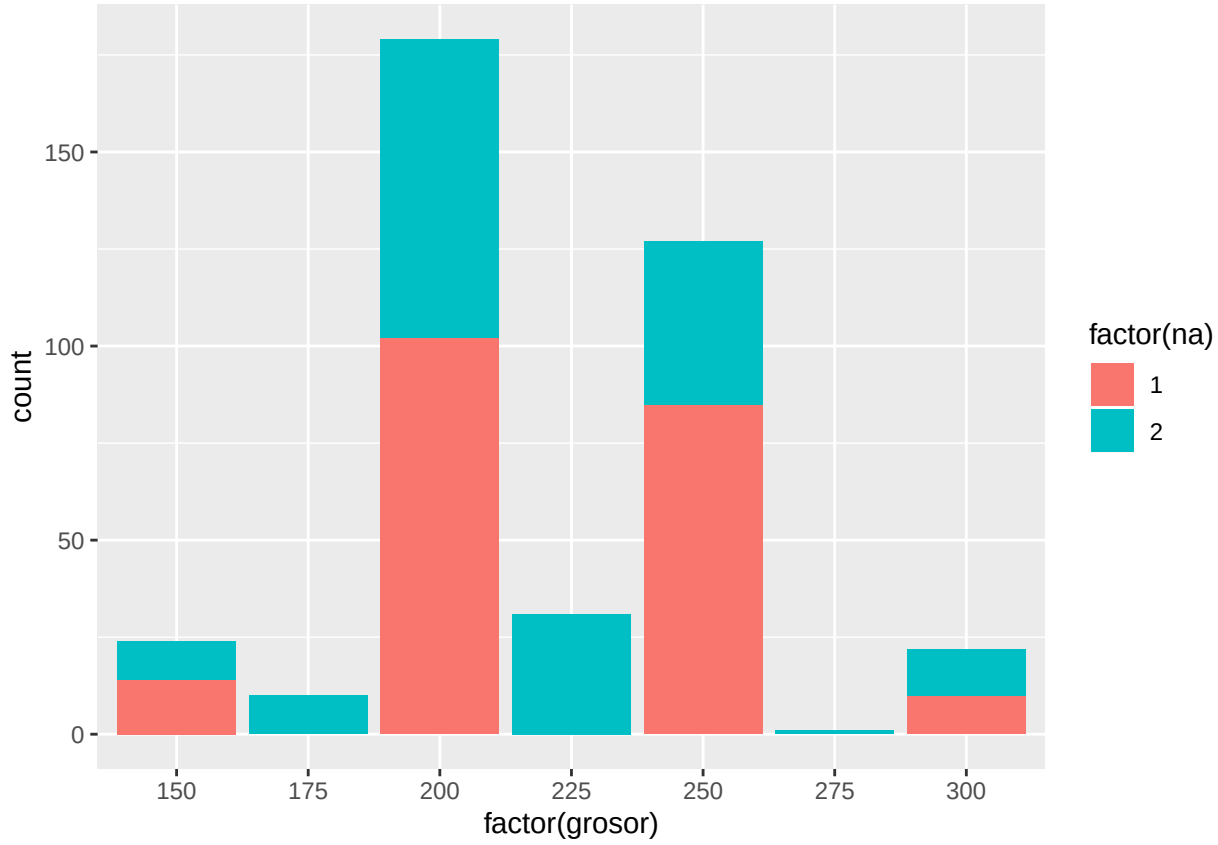
```
qplot(K1, K1.salida, data = queratocono) +
  geom_point() +
  xlab("K1") + ylab("K1.salida")
```



4. Build a histogram in terms of grosor (note that grosor should be taken as a factor) of the inserted ring.

The way qplot color codes the bars is opposite to how the colors are displayed in the legend. It can be resolve this two different ways; either reversing the legend or specify the direction of the levels when transforming the transmission (na) variable into a factor. Both align the legend color coding layout to the color coding of the stacked bars and also determine which color is top versus on the bottom.

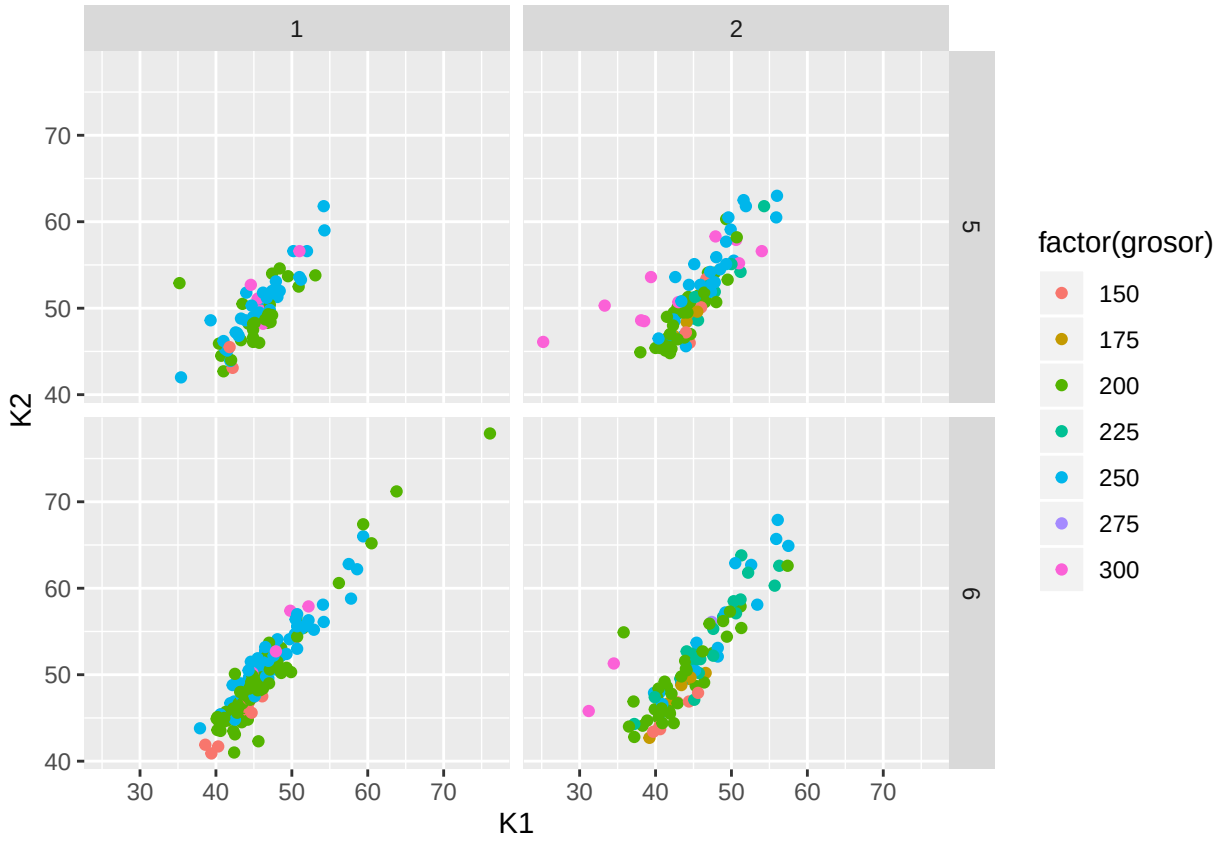
```
qplot(factor(grosor), data = queratocono, geom = "bar",
  fill = factor(na, levels = c(2, 1))) +
  guides(fill = guide_legend(reverse = TRUE)) +
  scale_fill_manual(values = c("#00BFC4", "#F8766D")) +
  labs(fill = "factor(na)") +
  ylab("count")
```



Manually the name of ylab has been changed since the method of qplot stat = "count" is deprecated. It is used by default but does not show the name in the graph, as the method of ggplot does.

5. Build a scatter plot of the relation between K1 and K2 with "faceting" in terms of the parameters diam and na, by assigning different colours to the points according to the thickness (grosor) of the ring. In order to visualise all points correctly use a transparency of value 1/3.

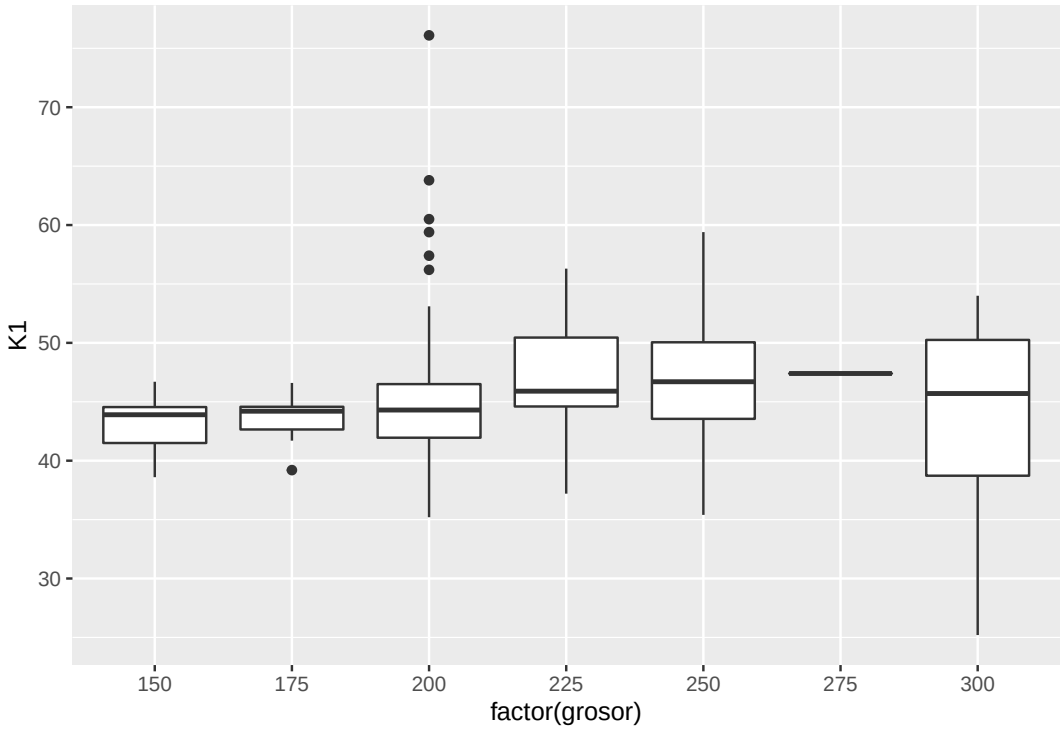
```
qplot(K1, K2, data = queratocono, colour = factor(grosor), facets = diam ~ na,
      size = I(1/3)) +
      geom_point() +
      scale_shape_manual(values = 0:7) +
      xlab("K1") + ylab("K2")
```

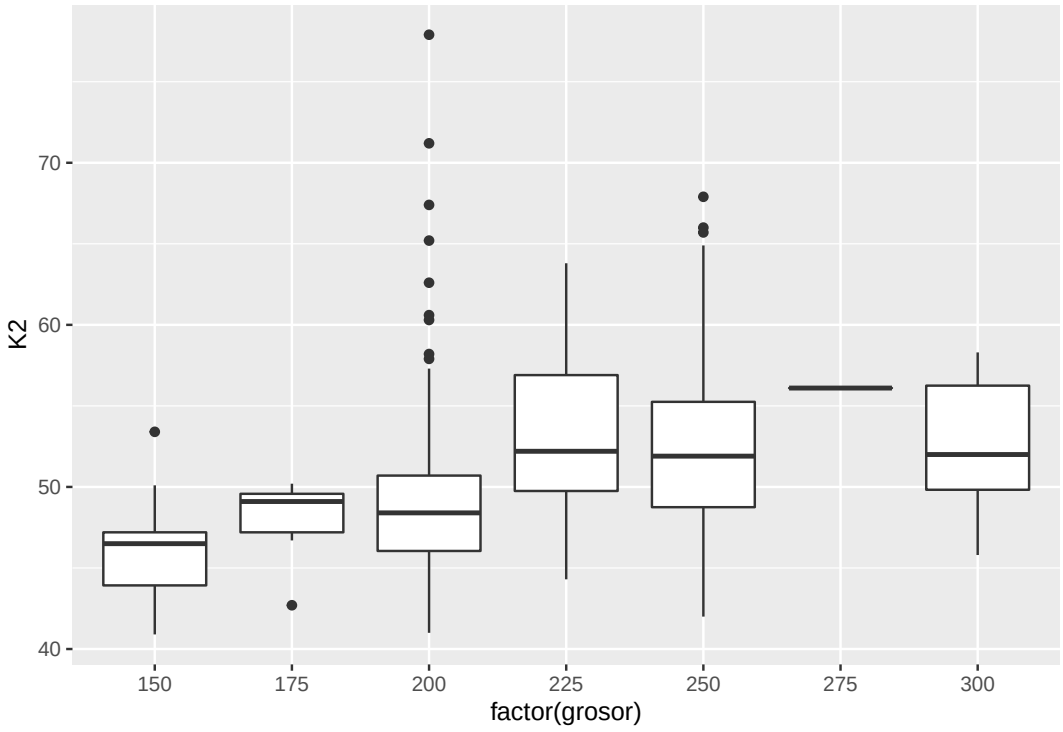


6. Create two boxplots that show a summary of the distributions of K1 and K2 (separately) with respect to the thickness (grosor).

```
qplot(factor(grosor), K1, data = queratocono, geom = "boxplot") +
  xlab("factor(grosor)") + ylab("K1")
```

```
qplot(factor(grosor), K2, data = queratocono, geom = "boxplot") +
  xlab("factor(grosor)") + ylab("K2")
```





Question 1. ¿Están K1 y K2 correlacionados? ¿Positiva o negativamente?

K1 y K2 están correlacionados porqué los puntos no están distribuidos uniformemente por toda la gráfica, si no que los puntos se disponen bastante alineados tanto en la curva como en la recta de regresión, aunque vemos que hay algunos outliers.

K1 y K2 están correlacionados positivamente porqué la pendiente de regresión es positiva, tanto en la curva como en la recta de regresión.

En este tipo de gráficos podemos observar que existe un problema de solapamiento y no podemos conocer realmente la dispersión de los puntos.

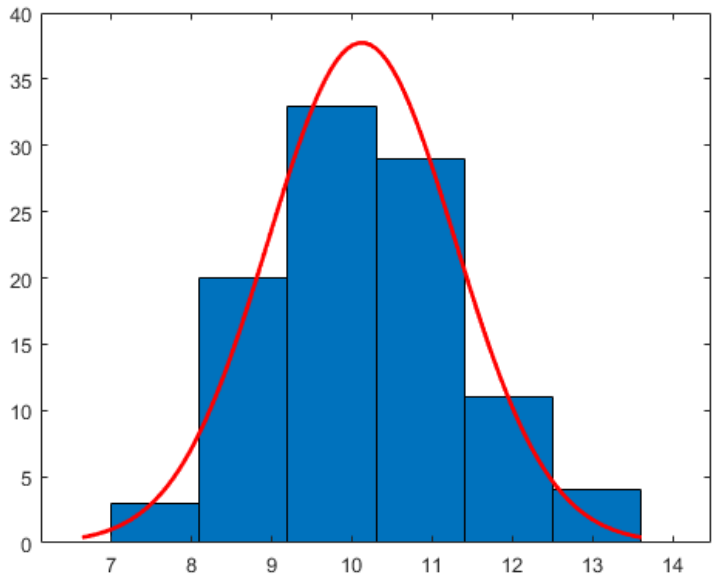
Question 2. Mirando a la figura 2 ¿vemos una correlación más fuerte para na=1? ¿Se debe a que la pendiente es ligeramente mayor?

La correlación es más fuerte para na=1 pero no se debe a que la pendiente sea ligeramente mayor. Se debe a que la recta de regresión es más alargada. Que la nube de puntos sea más estrecha y alargada indica que la correlación es más fuerte.

Mirando a la figura 2, también podemos observar que existe un problema de solapamiento y no podemos conocer realmente la dispersión de los puntos.

Question 3. Mirando el histograma, ¿es la distribución del grosor normal? Justifica la respuesta.

Mirando el histograma, la distribución del grosor no es normal. Ya que una distribución normal es un histograma idealizado, liso, en forma de campana, con toda la aleatoriedad eliminada, que representa un conjunto de datos ideal que tiene una gran cantidad de números concentrados en el medio del rango, con el resto de los números simétricamente en ambos lados. Por ejemplo,



En cambio, en nuestro histograma las barras son alternativamente altas y cortas, siguiendo una Comb Distribution. Este tipo de distribución a menudo resulta de datos redondeados y/o de un histograma construido incorrectamente.

Question 4. Mirando a los diagramas de bigotes, ¿qué ocurre con la caja para el grosor 275?

Observamos que la caja para el grosor 275 no existe. No se ha representado el rango intercuartílico (RIC), la diferencia entre el valor del tercer cuartil y el primer cuartil.

Investigando un poco la base de datos, vemos que solo existe un único valor para el grosor 275. La ausencia de datos podría ser la causa de que no exista caja para el grosor 275.