Graficos with ggplot

Agustin Muñoz Gonzalez

11/6/2020

Preparamos el entorno.

Limpiamos los registros y attachamos. (setear el directorio de trabajo!)

```
rm(list=ls())
tit=read.csv('titanic.csv', header=T)
attach(tit)
library(ggplot2)
```

¿Cuándo usamos cada gráfico?

Antes que nada mencionar que no todos los tipos de gráficos sirven para lo mismo.

Por ejemplo si queremos ver relaciones entre variables de tipo numérico, un grafico de barras no va a aportar información muy clara, y en cambio un gráfico de dispersión es mas adecuado. Pero si nos interesa ver como se distribuyen los datos de cierta variable numerica en otra variabla categórica entonces el adecuado sería un gráfico de caja, ya que un diagrama de dispersión, de barras o de torta serían simplemente manchas negras (si los datos numéricos son muchos).

Como para tener en la cabeza, los distintos gráficos son adecuados para las siguientes situaciones (a grandes rasgos)

- Histograma: Para estudiar densidad de una variable numérica.
- Gráfico de caja: Para estudiar var categorica vs var numérica. Es decir, la disposición de la numérica en las categorías o grupos de la categórica.
- Diagrama de dispersión: Para estudiar var numérica vs var numérica.
- Gráfico de barras o de torta: Para estudiar la distribución de una (o más) var categórica en el total de datos.

Scatterplots

Me basé en https://mathstat.slu.edu/~speegle/_book/ggplot.html

Carguemos el set de datos mpg de la libreria ggplot2.

```
data(mpg)
head(mpg)
```

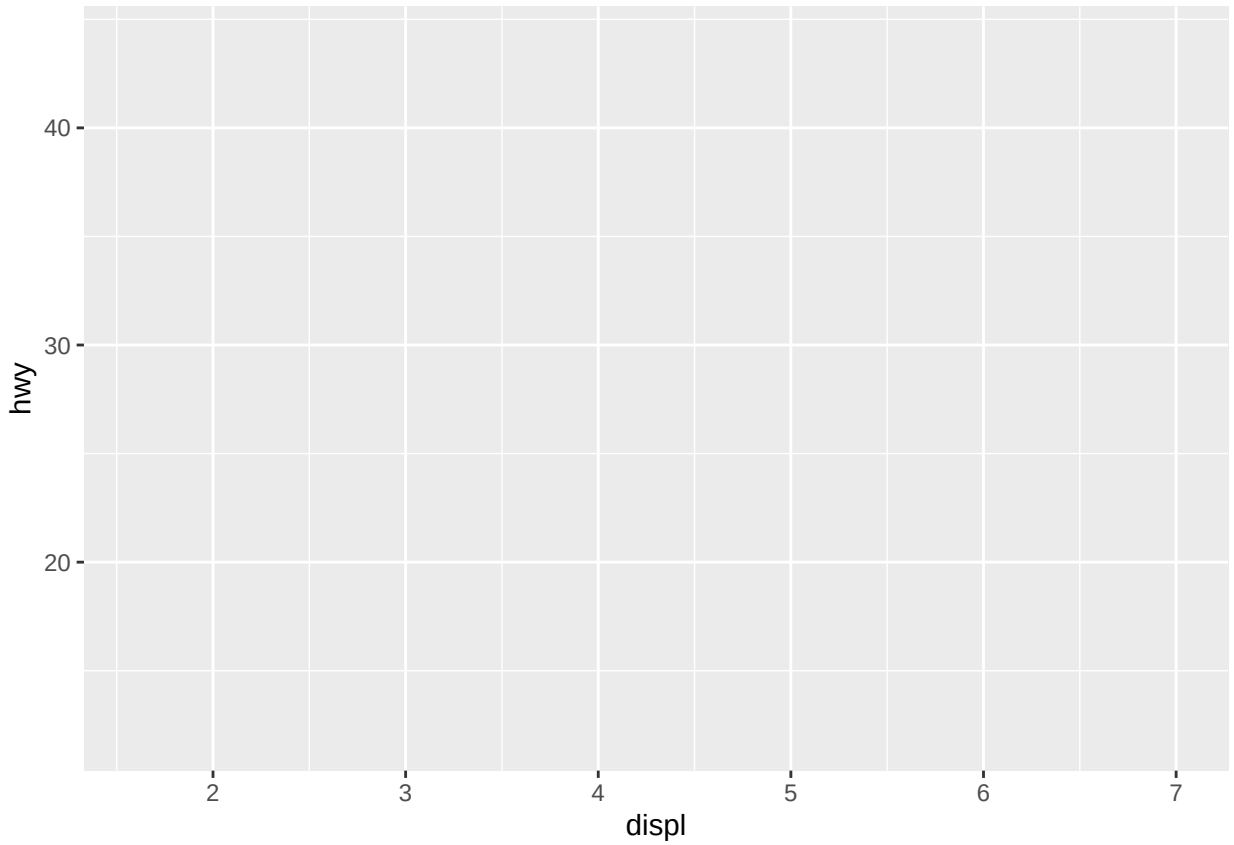
```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

```
tail(mpg)
```

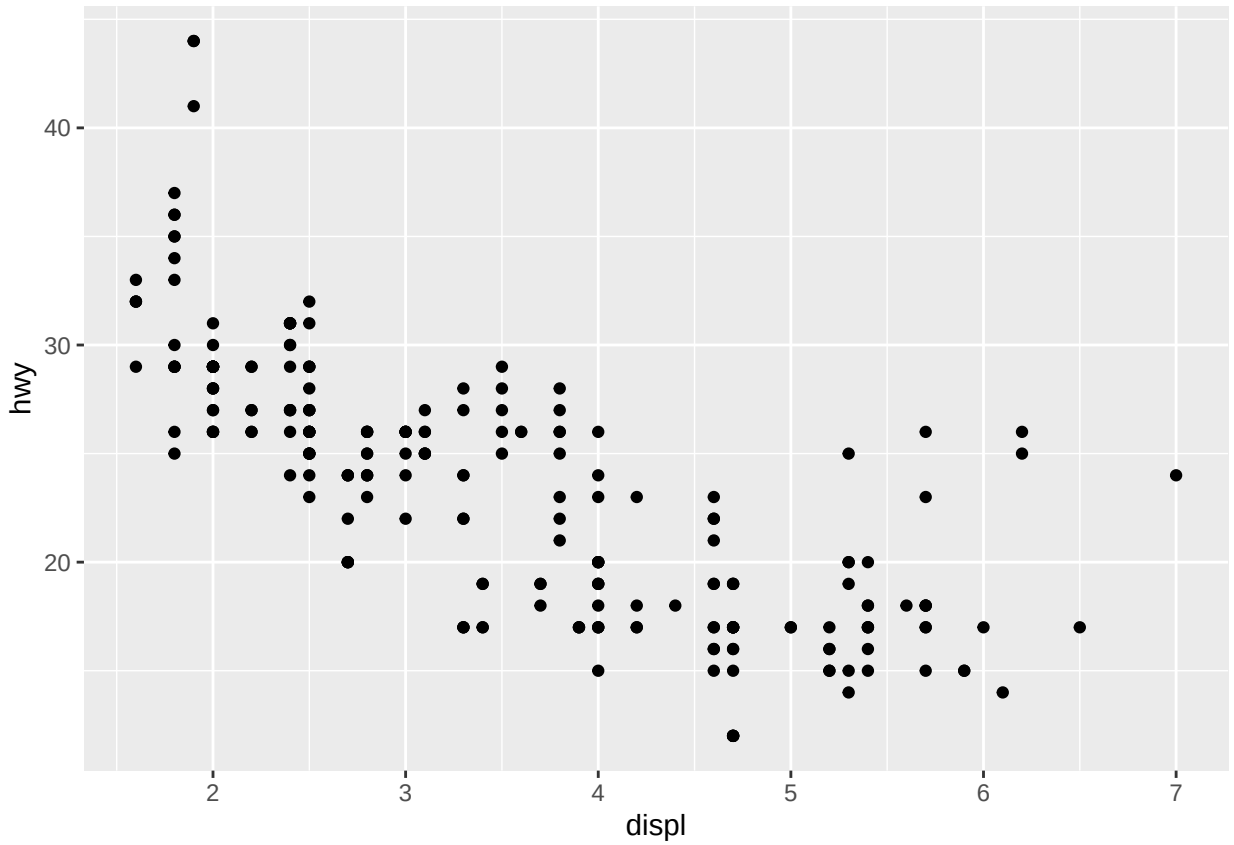
```
## # A tibble: 6 x 11
##   manufacturer model  displ  year   cyl trans     drv     cty   hwy fl    class
##   <chr>        <chr>  <dbl> <int> <int> <chr>     <chr> <int> <int> <chr> <chr>
## 1 volkswagen   passat   1.8  1999     4 auto(l5)  f        18    29 p     midsi~
## 2 volkswagen   passat   2    2008     4 auto(s6)  f        19    28 p     midsi~
## 3 volkswagen   passat   2    2008     4 manual(m~ f        21    29 p     midsi~
## 4 volkswagen   passat   2.8  1999     6 auto(l5)  f        16    26 p     midsi~
## 5 volkswagen   passat   2.8  1999     6 manual(m~ f        18    26 p     midsi~
## 6 volkswagen   passat   3.6  2008     6 auto(s6)  f        17    26 p     midsi~
```

Grafiquemos un poco.

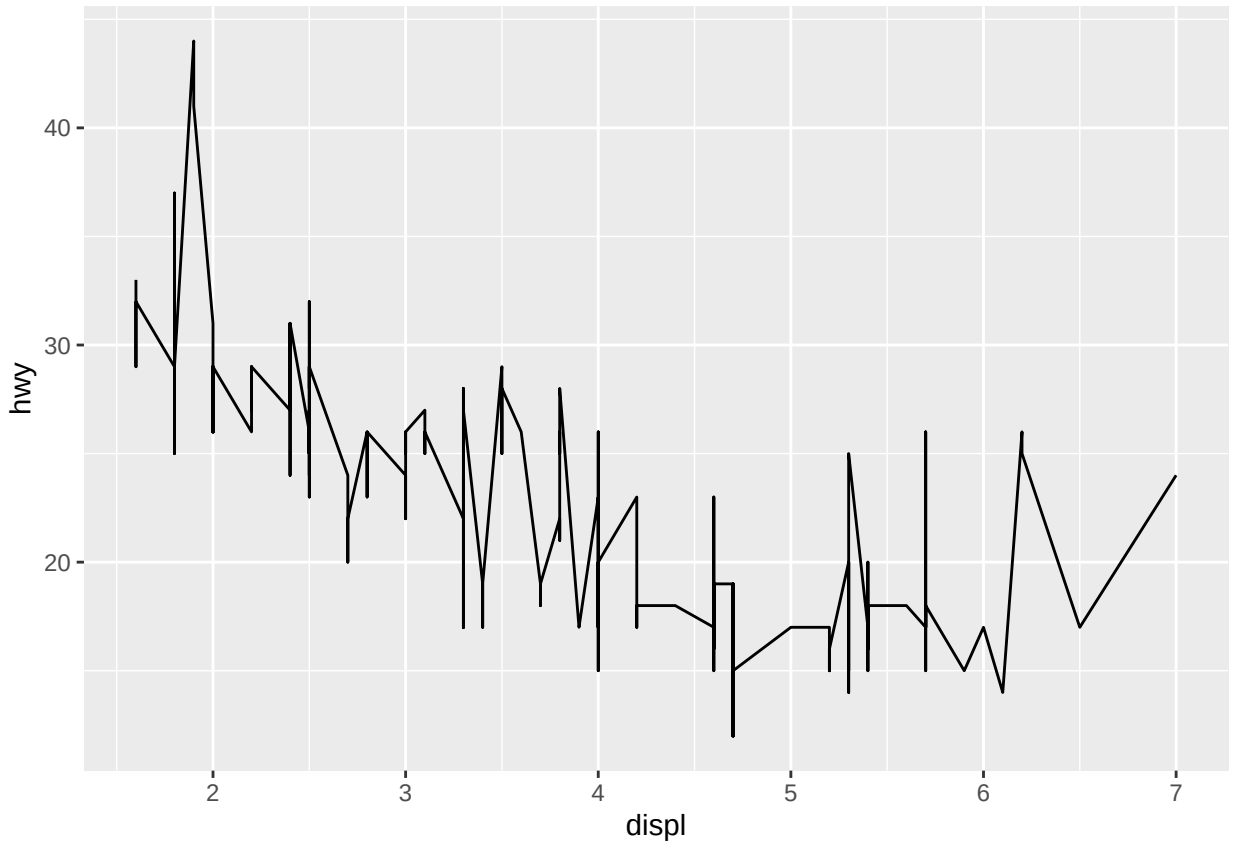
```
ggplot(data = mpg, aes(x=displ, y=hwy))
```




```
ggplot(data = mpg, aes(x=displ, y=hwy))+
  geom_point()
```

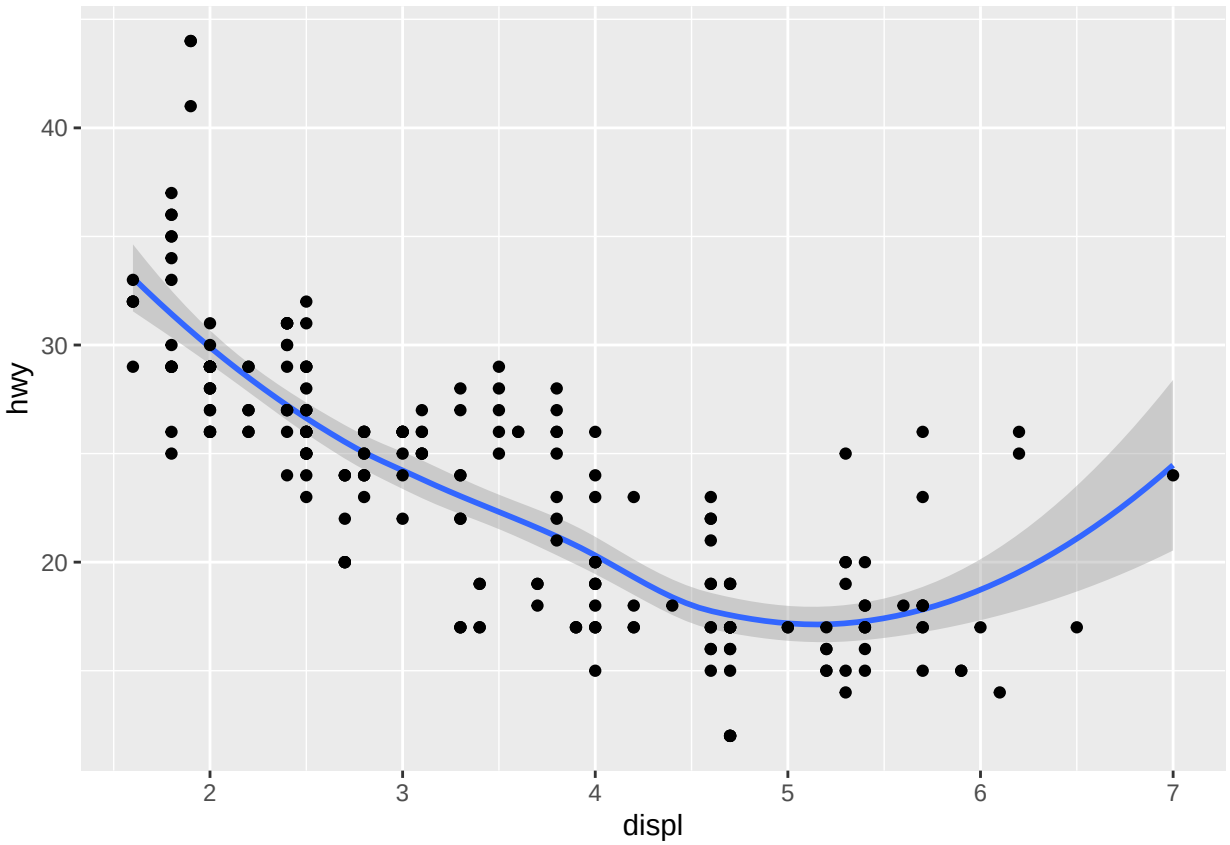



```
ggplot(data = mpg, aes(x=displ, y=hwy))+
  geom_line()
```



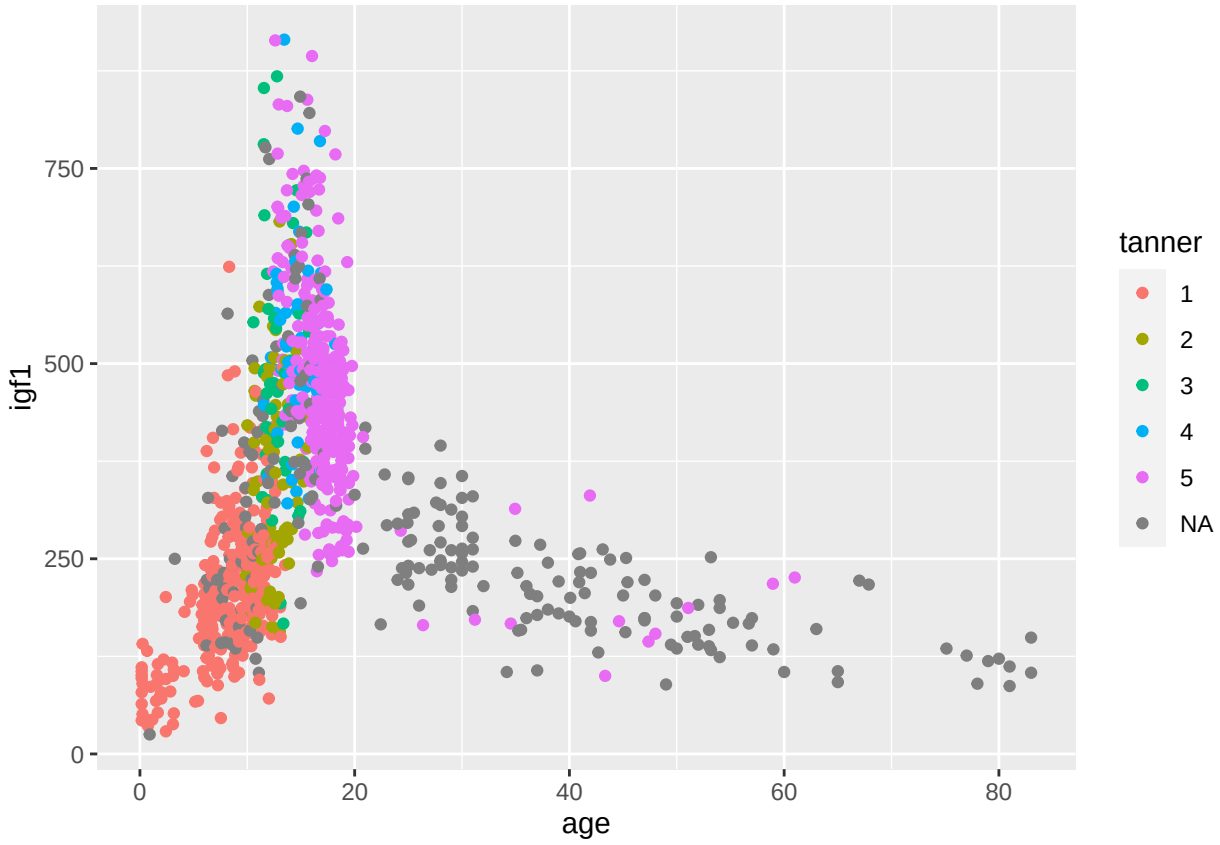

```
ggplot(data = mpg, aes(x=displ, y=hwy))+
  geom_smooth()+
  geom_point()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


40
30
20
hwy
2
3
4
5
6
7
displ

```
library(ISwR)
juul$tanner=as.factor(juul$tanner)
juul$sex=as.factor(juul$sex)
ggplot(juul, aes(x = age,y = igf1, color = tanner)) +
  geom_point()
```

```
## Warning: Removed 326 rows containing missing values (geom_point).
```

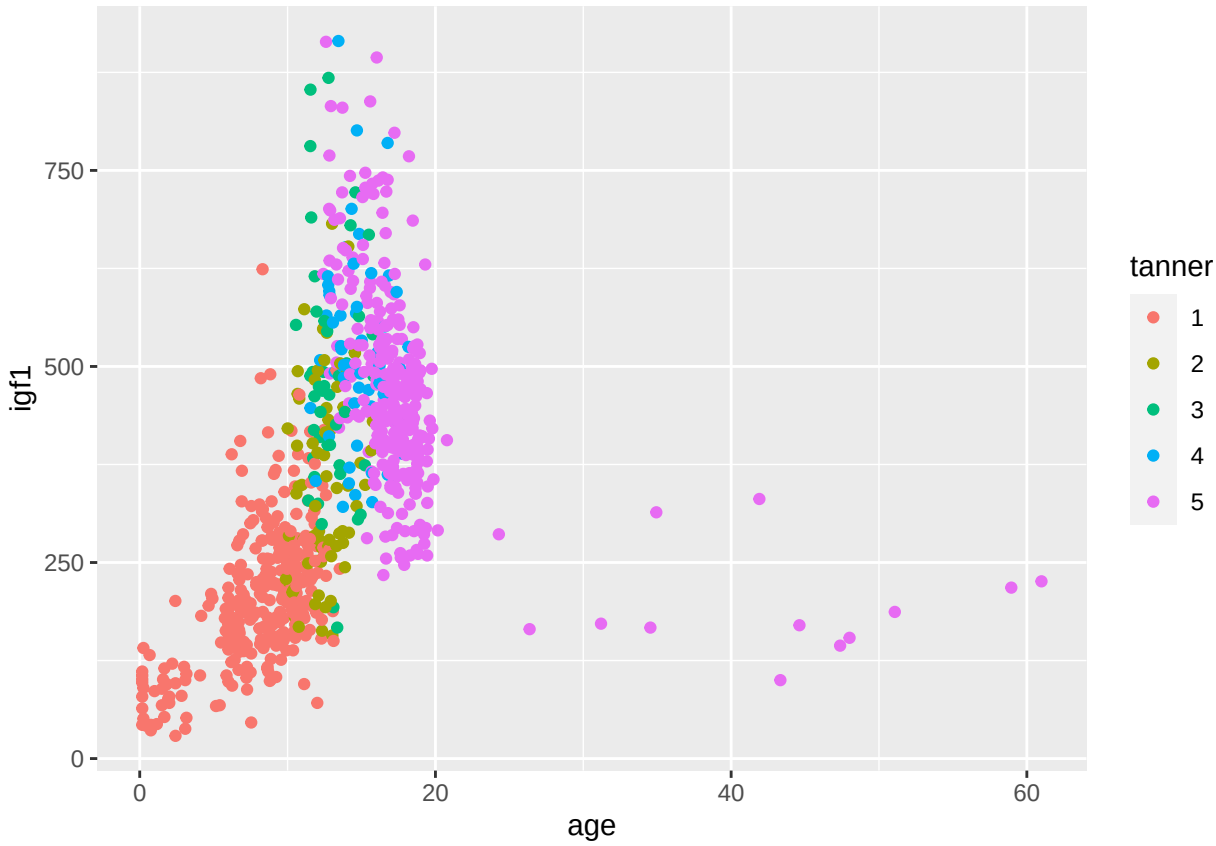



```
  ggtitle('Age vs igf1')
```

```
## $title
## [1] "Age vs igf1"
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(juul[!is.na(juul$tanner),], aes(x = age,y = igf1, color = tanner)) +
  geom_point()
```

```
## Warning: Removed 307 rows containing missing values (geom_point).
```

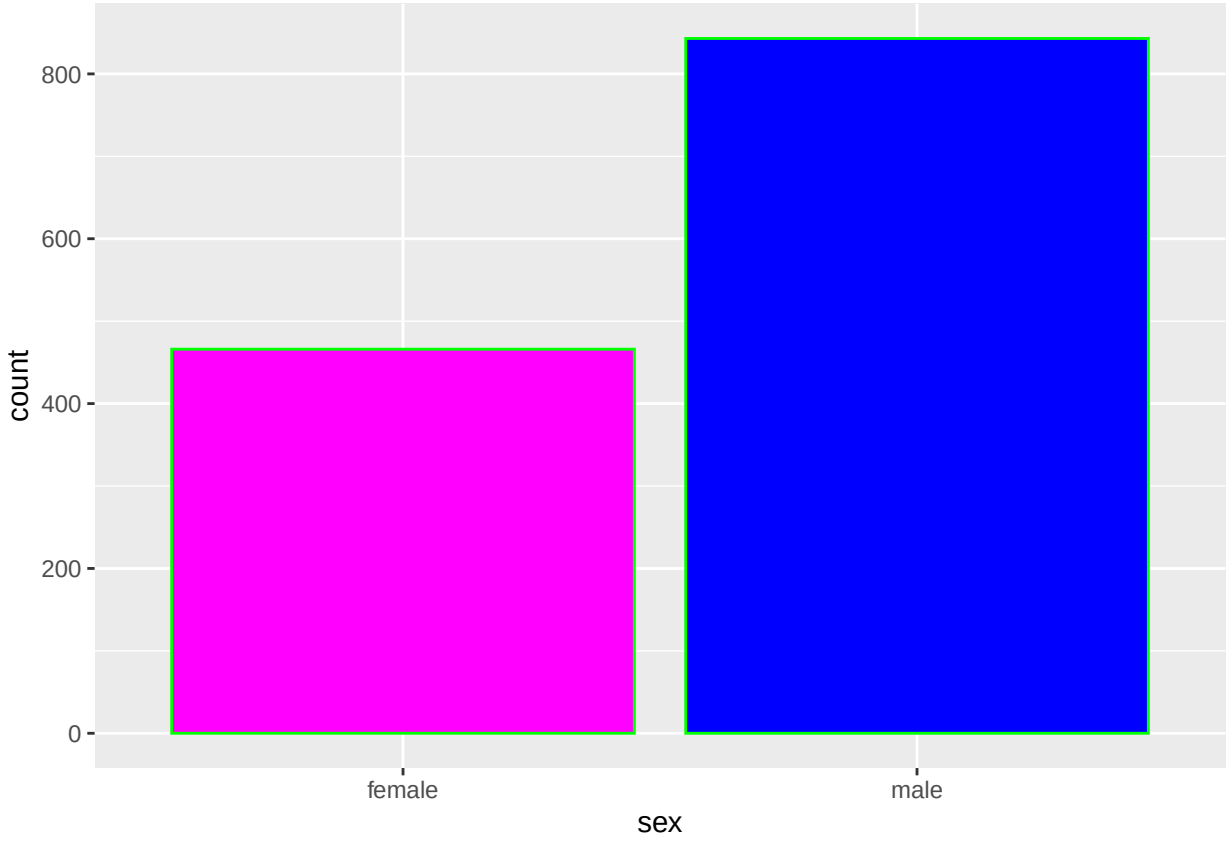



```
  labs(title ="Age vs igf1", x = "Age", y = "IGF1")
```

```
## $x
## [1] "Age"
##
## $y
## [1] "IGF1"
##
## $title
## [1] "Age vs igf1"
##
## attr(,"class")
## [1] "labels"
```

Barplots

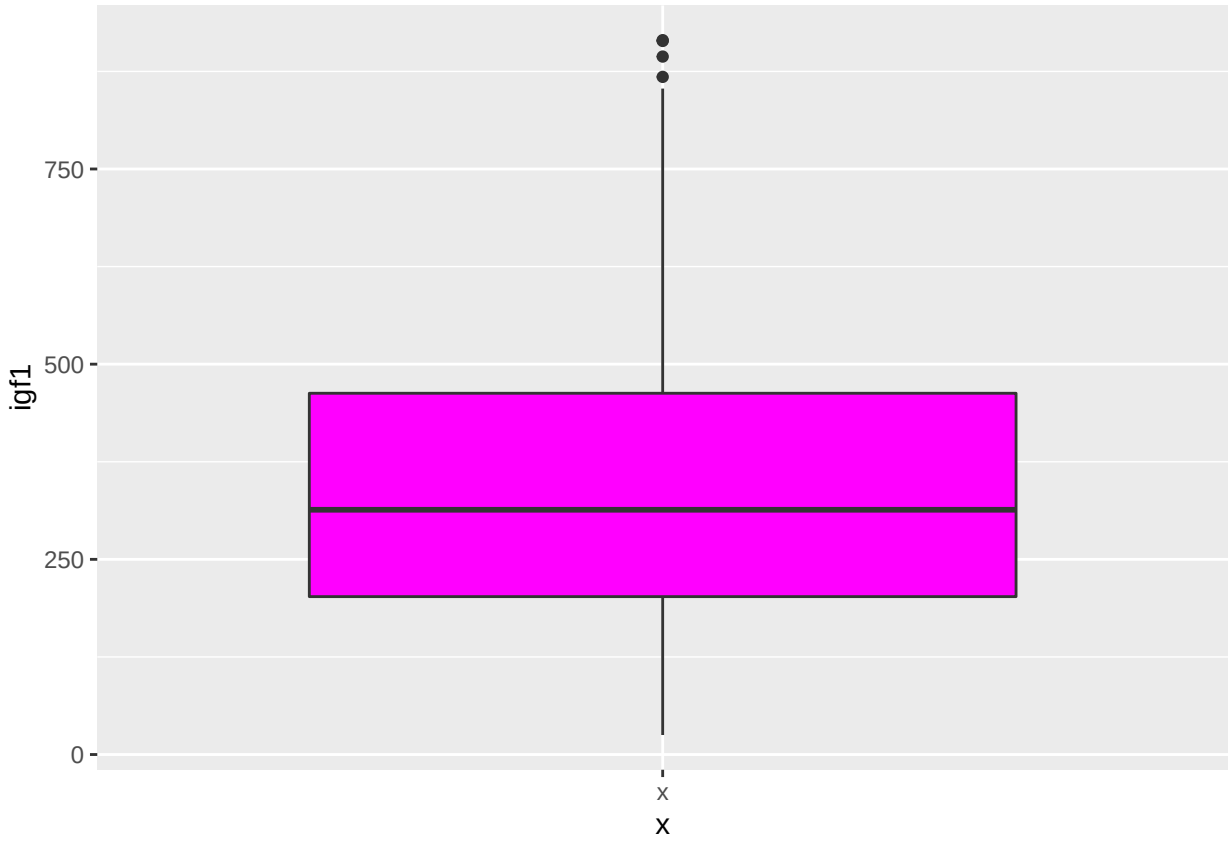
```
ggplot(tit,aes(x=sex))+
  geom_bar(color='green', fill=c('magenta','blue'))
```



Boxplots

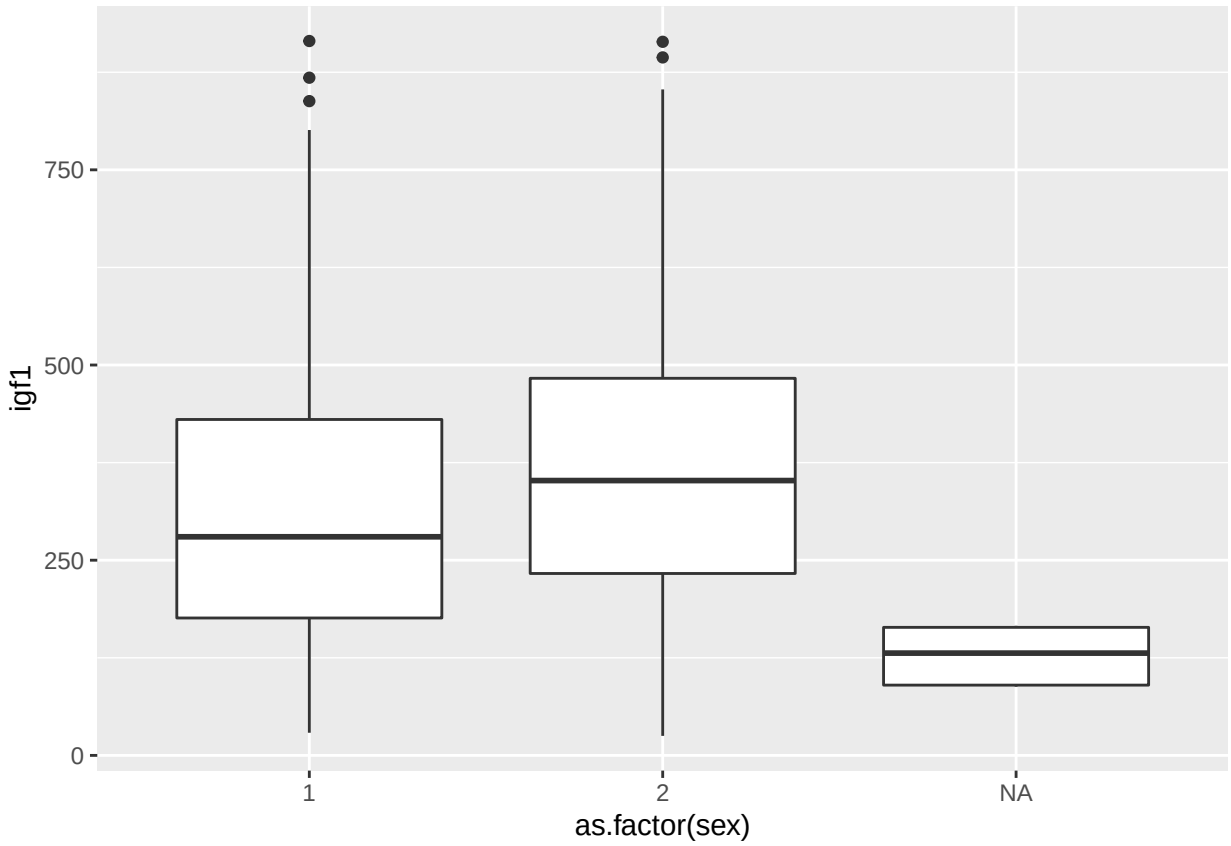
```
ggplot(juul, aes(x = "x", y = igf1)) +
  geom_boxplot(fill='magenta')
```

```
## Warning: Removed 321 rows containing non-finite values (stat_boxplot).
```



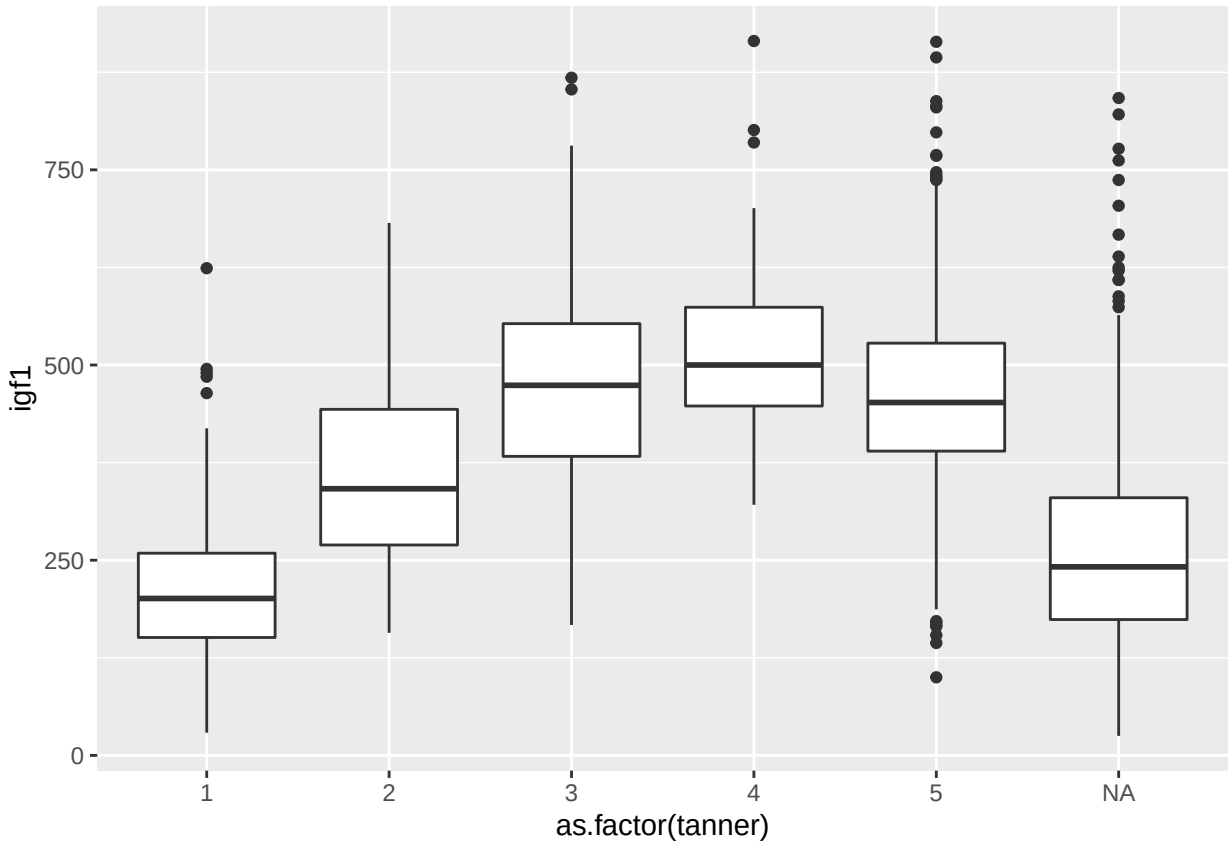

```
ggplot(juul, aes(x = as.factor(sex), y = igf1)) +
  geom_boxplot()
```

```
## Warning: Removed 321 rows containing non-finite values (stat_boxplot).
```




```
# no anda: era porque sex no era un factor!
ggplot(juul, aes(x = as.factor(tanner), y = igf1)) +
  geom_boxplot()
```

```
## Warning: Removed 321 rows containing non-finite values (stat_boxplot).
```

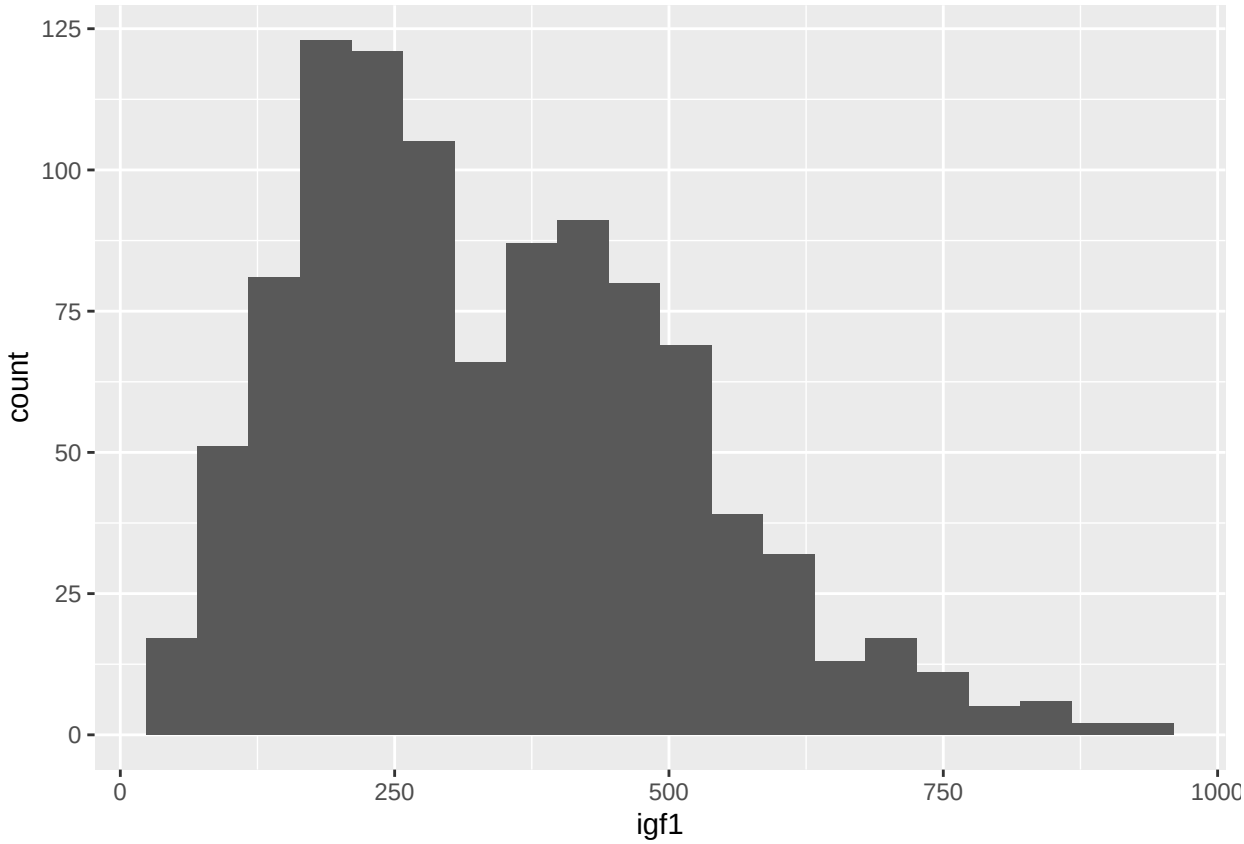



```
# no anda: era porque tanner no era un factor!
```

Histograms

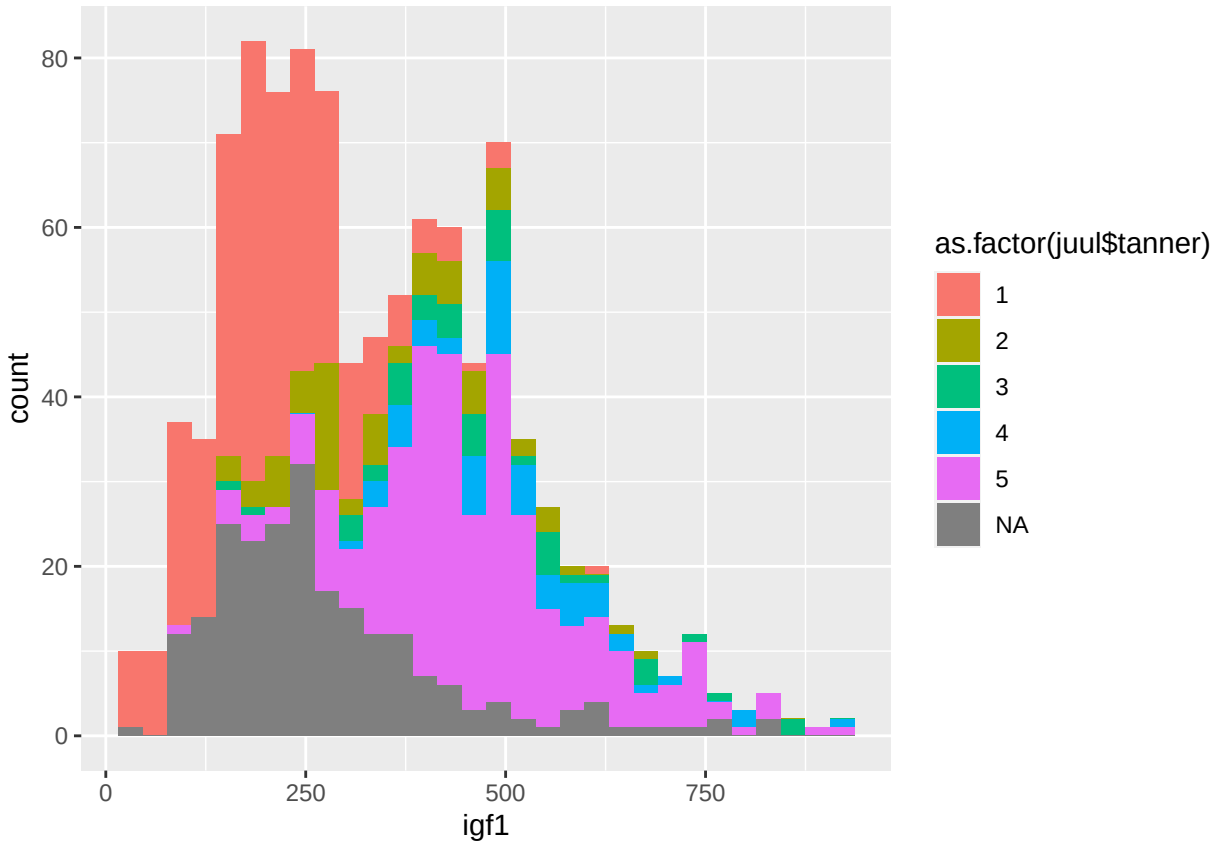
```
ggplot(juul, aes(x = igf1)) +
  geom_histogram(bins = 20)
```

```
## Warning: Removed 321 rows containing non-finite values (stat_bin).
```




```
#bins = cantidad decolumnas del histograma
ggplot(juul, aes(x = igf1)) +
  geom_histogram(aes(fill = as.factor(juul$tanner)))
```

```
## Warning: Use of `juul$tanner` is discouraged. Use `tanner` instead.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 321 rows containing non-finite values (stat_bin).
```




```
# Por qué no los pinta? será que el largo de tanner
# no es el mismo que la cantidad de columnas (=bins)?
# No, era porque el fill tiene que ser un factor!
```

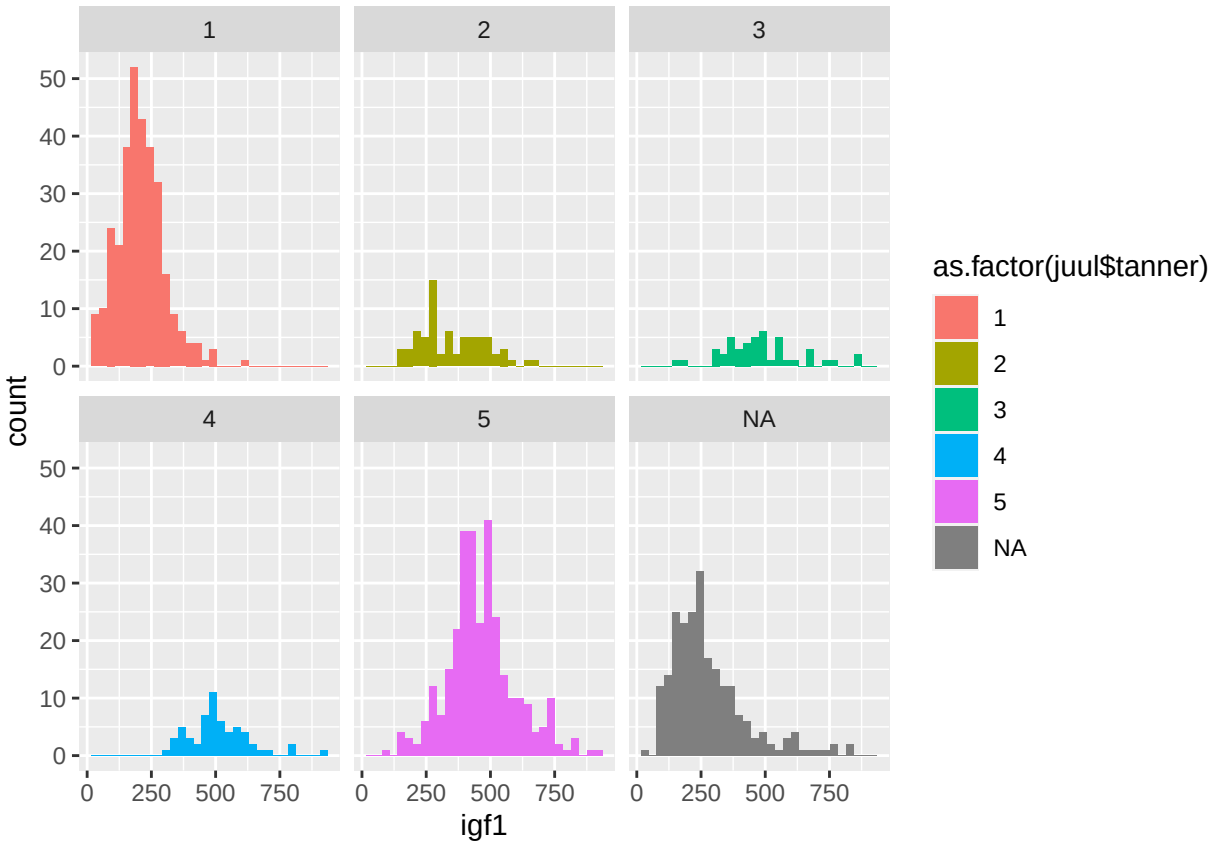
Better would be if we had each histogram plotted on its own axes. In the graphical lexicon, that is called faceting.

```
ggplot(juul, aes(x = igf1)) +
  geom_histogram(aes(fill = as.factor(juul$tanner))) +
  facet_wrap(~tanner)
```

```
## Warning: Use of `juul$tanner` is discouraged. Use `tanner` instead.
```

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

```
## Warning: Removed 321 rows containing non-finite values (stat_bin).
```



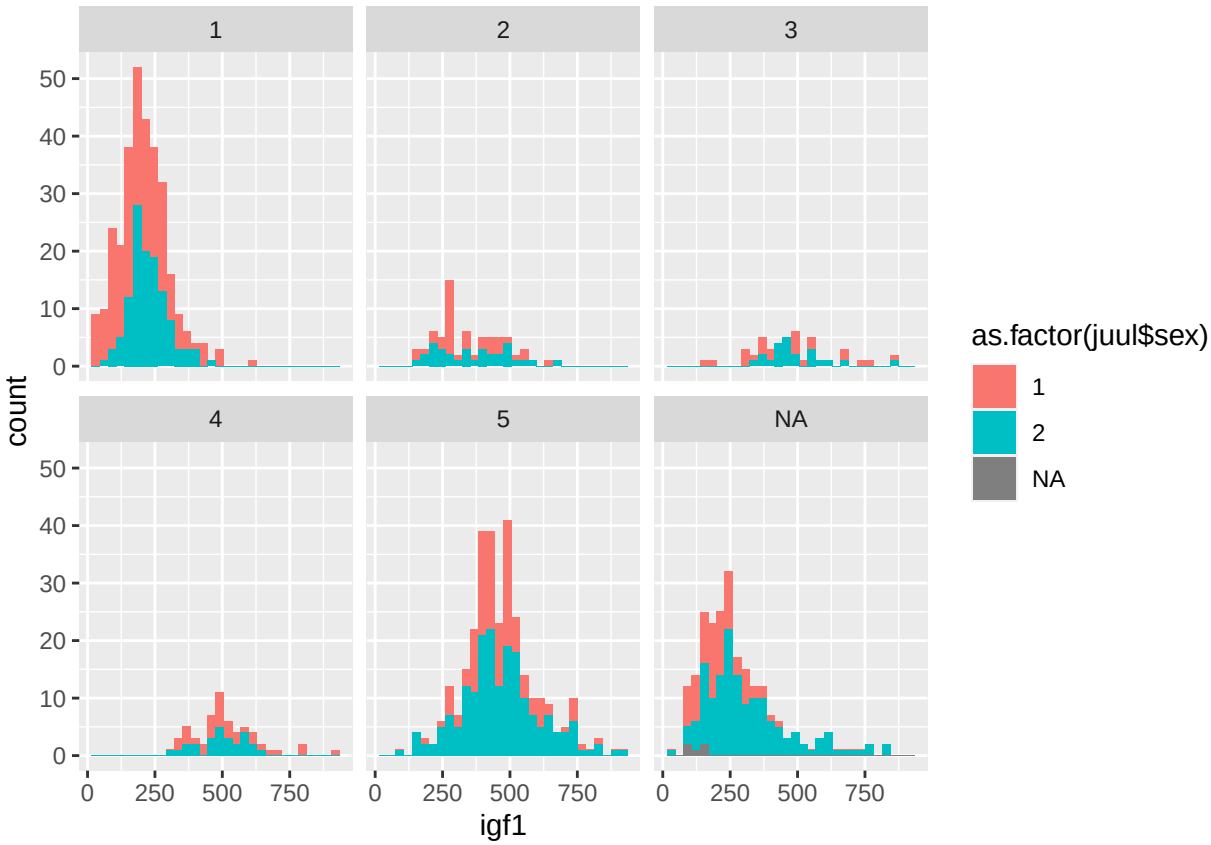
Now, though, the colors don't mean anything, so we could color by another variable instead of tanner, such as by sex.

```
ggplot(juul, aes(x = igf1)) +
  geom_histogram(mapping = aes(fill = as.factor(juul$sex))) +
  facet_wrap(~tanner)
```

```
## Warning: Use of `juul$sex` is discouraged. Use `sex` instead.
```

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

```
## Warning: Removed 321 rows containing non-finite values (stat_bin).
```

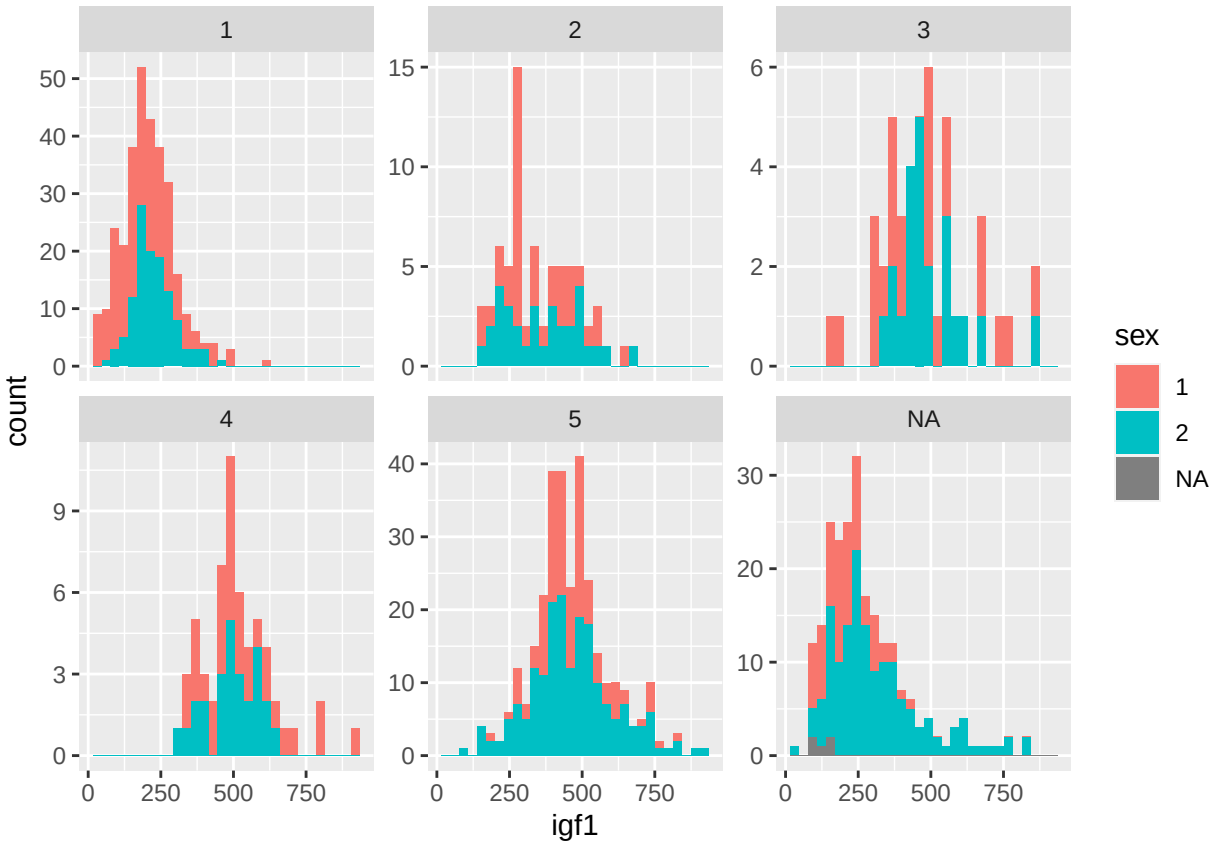


Occasionally, we prefer to free up the axes so that they are not all on the same scale.

```
ggplot(juul, aes(x = igf1)) +
  geom_histogram(mapping = aes(fill = sex)) +
  facet_wrap(~tanner, scales = "free_y")
```

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

```
## Warning: Removed 321 rows containing non-finite values (stat_bin).
```


1
2
3
4
5
NA
count
igf1
sex
1
2
NA