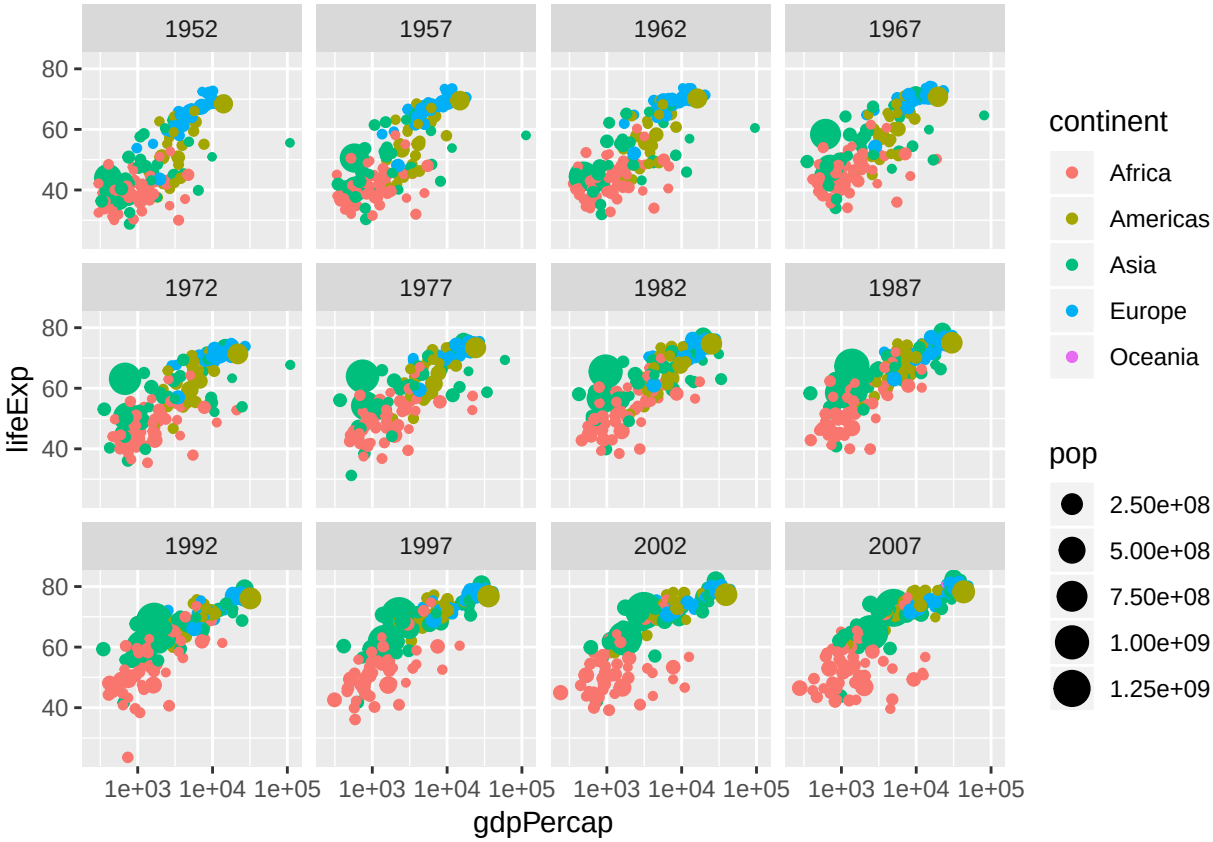
HW07

Oksana Ivanova

```
set.seed(42)
library(gapminder)
library(ggplot2)
library(datasets)
library(dplyr)
```

Data: Gapminder dataset, All years facet

```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, color = continent, size = pop)) +
  geom_point() +
  scale_x_log10() +
  facet_wrap(~ year)
```



Data: Airquality, transform, plot all measures by time

```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
airquality$Day = factor(airquality$Day)
airquality$Month = factor(airquality$Month)
```

```
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : Factor w/ 5 levels "5","6","7","8",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Day    : Factor w/ 31 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp       Month
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   5:31
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   6:30
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   7:31
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   8:31
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   9:30
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##       Day
##  1      :  5
##  2      :  5
##  3      :  5
##  4      :  5
##  5      :  5
##  6      :  5
##  (Other):123
```

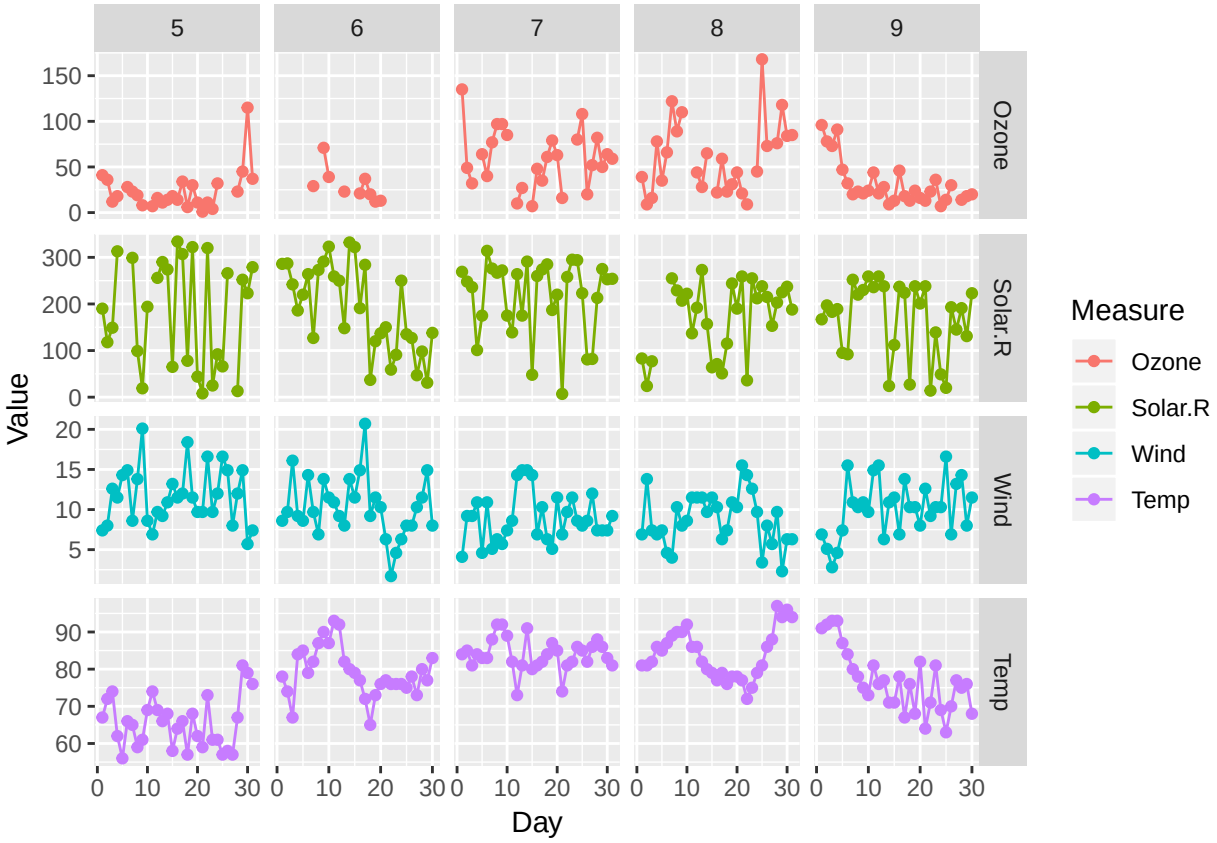
```
#Remove NA values
library(reshape2)
aqLong = melt(airquality, id.vars=c("Month", "Day"), variable.name = "Measure", value.name="Value")
aqLong$Measure = as.factor(aqLong$Measure)
aqLong$Day = as.numeric(aqLong$Day)
head(aqLong)
```

```
##   Month Day Measure Value
## 1     5   1   Ozone    41
## 2     5   2   Ozone    36
## 3     5   3   Ozone    12
## 4     5   4   Ozone    18
## 5     5   5   Ozone    NA
## 6     5   6   Ozone    28
```

View(aqLong)

```
ggplot(aqLong, aes(x = Day, y = Value, fill = Measure, colour = Measure)) +
  geom_point(aes(x = Day, y = Value)) +
  geom_line(aes(x = Day, y = Value)) +
  facet_grid(Measure ~ Month, scales = "free") +
  scale_x_continuous(breaks = seq(0, 31, by = 10))
```

```
## Warning: Removed 44 rows containing missing values (geom_point).
```



Some numeric data: distribution plots

```
data("diamonds")
df = diamonds
head(df)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
```

```
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

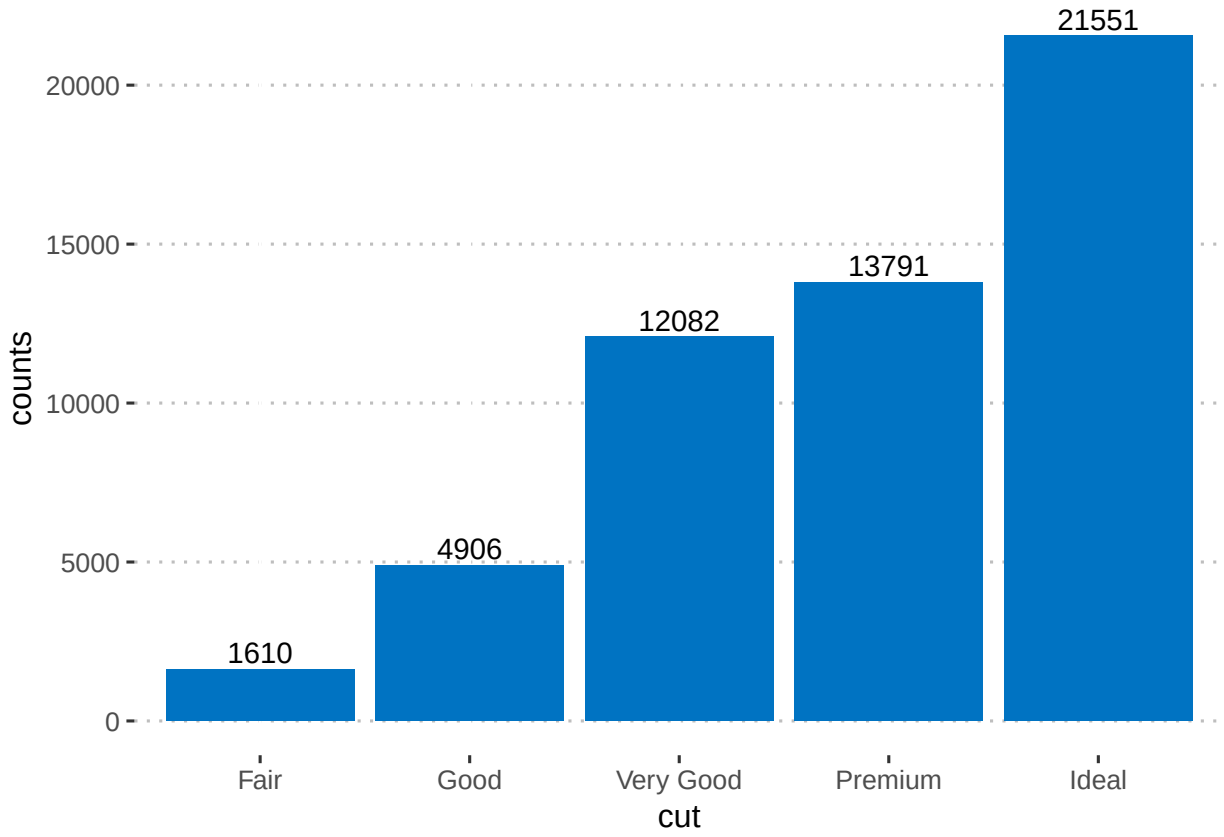
```
library(dplyr)
df <- diamonds %>%
  group_by(cut) %>%
  summarise(counts = n())
df
```

```
## # A tibble: 5 x 2
##   cut       counts
##   <ord>      <int>
## 1 Fair        1610
## 2 Good        4906
## 3 Very Good  12082
## 4 Premium    13791
## 5 Ideal      21551
```

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
ggplot(df, aes(x = cut, y = counts)) +
  geom_bar(fill = "#0073C2FF", stat = "identity") +
  geom_text(aes(label = counts), vjust = -0.3) +
  theme_pubclean()
```




```
df = ChickWeight
head(df)
```

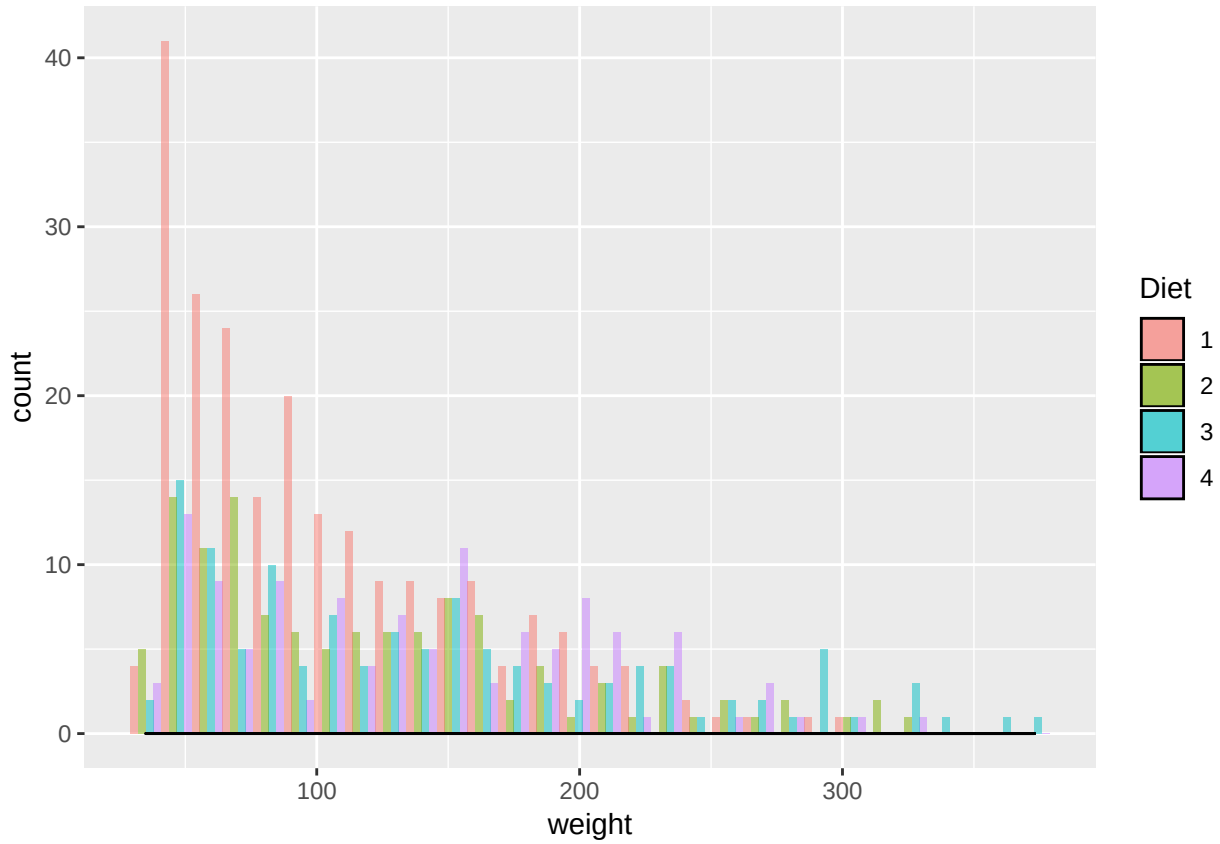
```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

```
ggplot(df, aes(x = weight, fill = Diet)) +
  geom_histogram()
```

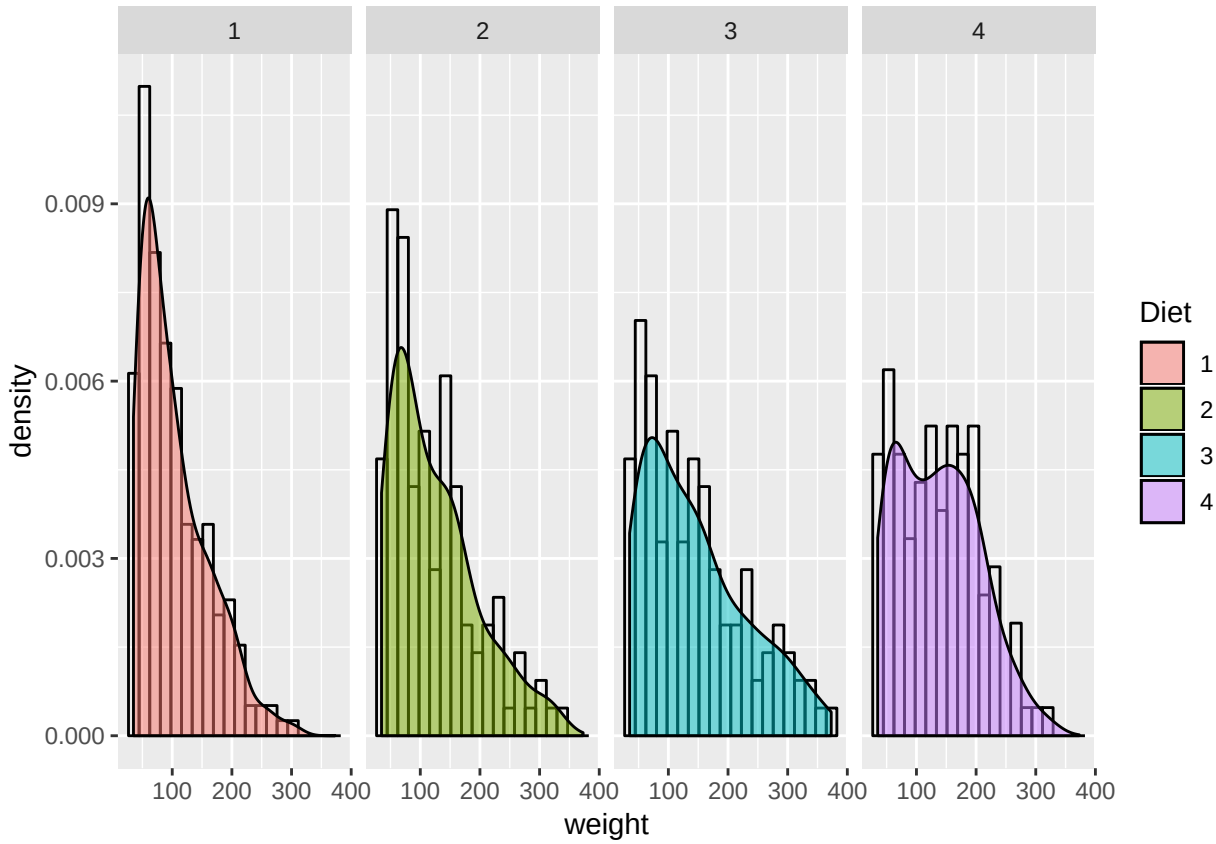
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


```
ggplot(df, aes(x = weight, fill = Diet)) +
  geom_histogram(alpha =.5, position = "dodge") +
  geom_density(alpha = 0.3)
```

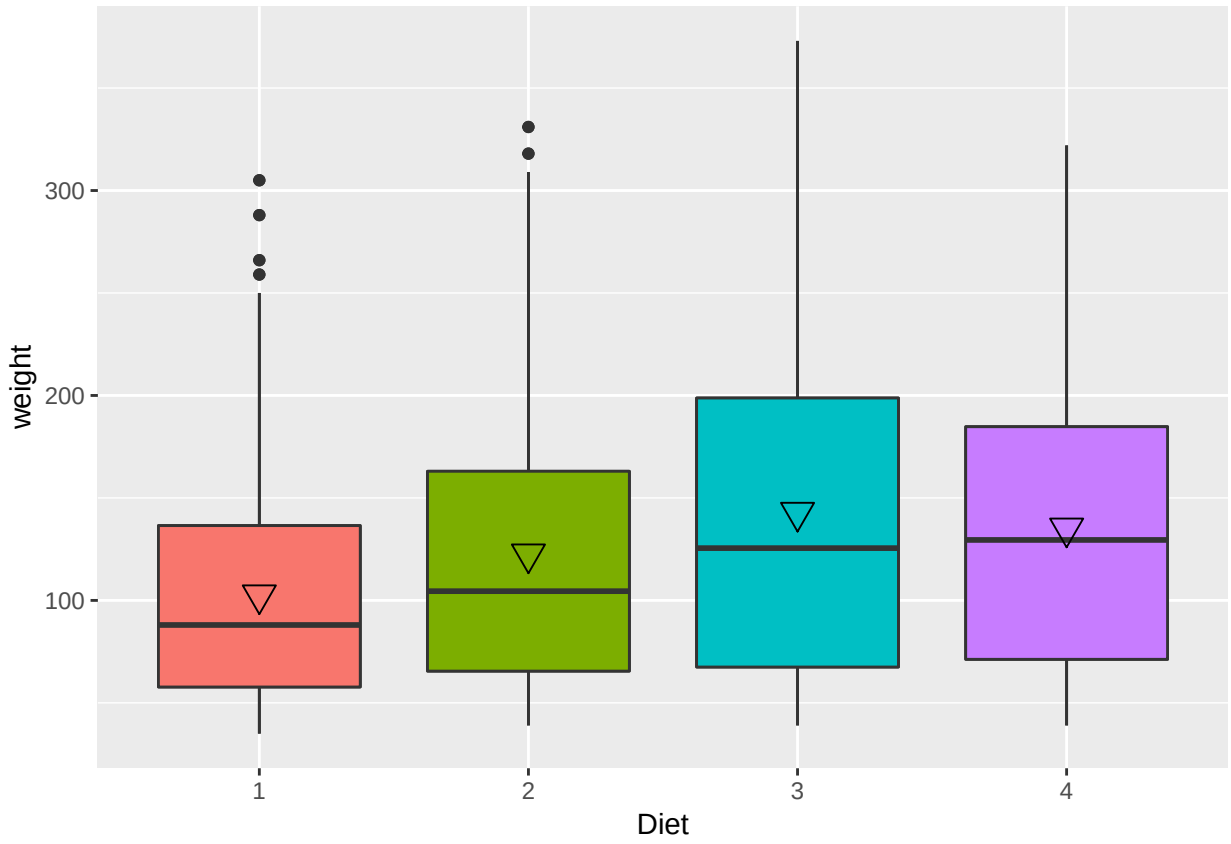
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(x = weight, fill = Diet)) +
  geom_histogram(aes(y = ..density..), bins = 20, position = "identity", alpha = 0, color = "black") +
  geom_density(alpha = 0.5) +
  facet_grid(.~ Diet)
```




```
ggplot(df, aes(x = Diet, y = weight, fill = Diet)) +
  geom_boxplot() +
  guides(fill = FALSE) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 6, size = 4)
```


weight
300
200
100
1
2
3
4
Diet