Introduction to ggplot2

Karen Mazidi

Code Accompanying The Machine Learning Handbooks, Volume I, Chapter 4

This code demonstrates the ggplot package.

Packages needed in this notebook:

```
if (!require(tidyverse)){
  install.packages("tidyverse")
}
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.0
## v purrr     1.0.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("tidyverse")

if (!require(mlbench)){
  insyes
  tall.packages("mlbench")
}
```

```
## Loading required package: mlbench
```

```
library("mlbench")

if (!require(gridExtra)){
  insyes
  tall.packages("gridExtra")
}
```

```
## Loading required package: gridExtra
##
## Attaching package: 'gridExtra'
##
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library("gridExtra")
```

There are 7 grammatical elements in ggplot2, the first 3 of these are essential to getting something plotted:

- data - the data being plotted should be the first argument, or specify data=...
- aesthetics - the scales onto which we plot; use aes() to specify at least x= and y= if needed as well as other parameters for customization
- geometries - visual elements such as points, lines, etc.
- facets - for plotting multiples
- statistics - representations to aid understanding
- coordinates - space on which data will be plotted
- themes - you can customize your own theme to use over and over

load tidyverse and some data

Loading the diabetes data set from package mlbench.

```
data("PimaIndiansDiabetes2")

tb <- as_tibble(PimaIndiansDiabetes2)
```

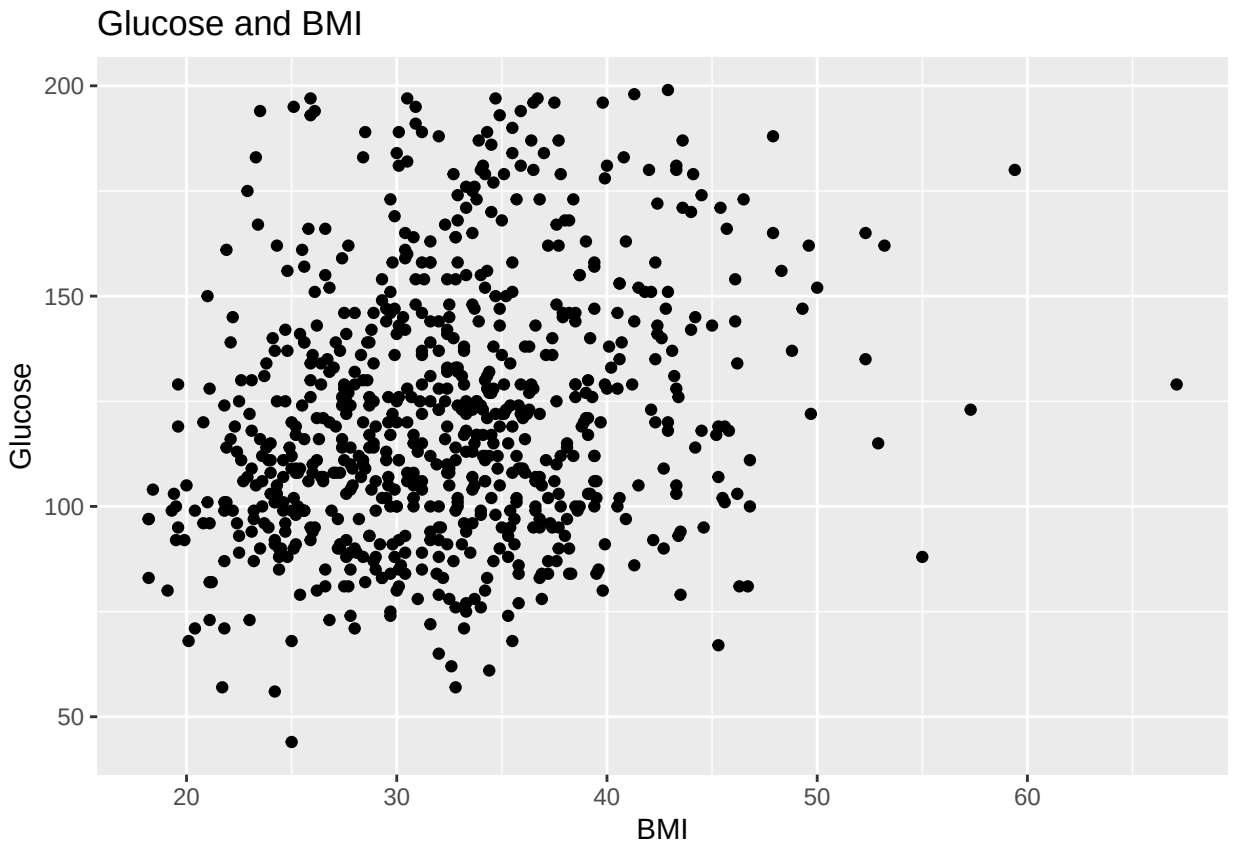
Explore ggplot2

Hadley Wickham developed ggplot2 in 2005, inspired by a grammar of graphics developed by Leland Wildinson in 1999. The ggplot2 functions are much more powerful than standard R graphs but also slower.

We have a short example below showing important components of building a ggplot. First we specify the data, then the aesthetics which are how the data is represented, followed by the geometry and finally labels.

```
ggplot(tb, aes(x=mass, y=glucose)) +
  geom_point() +
  labs(title="Glucose and BMI", x="BMI", y="Glucose")
```

```
## Warning: Removed 16 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



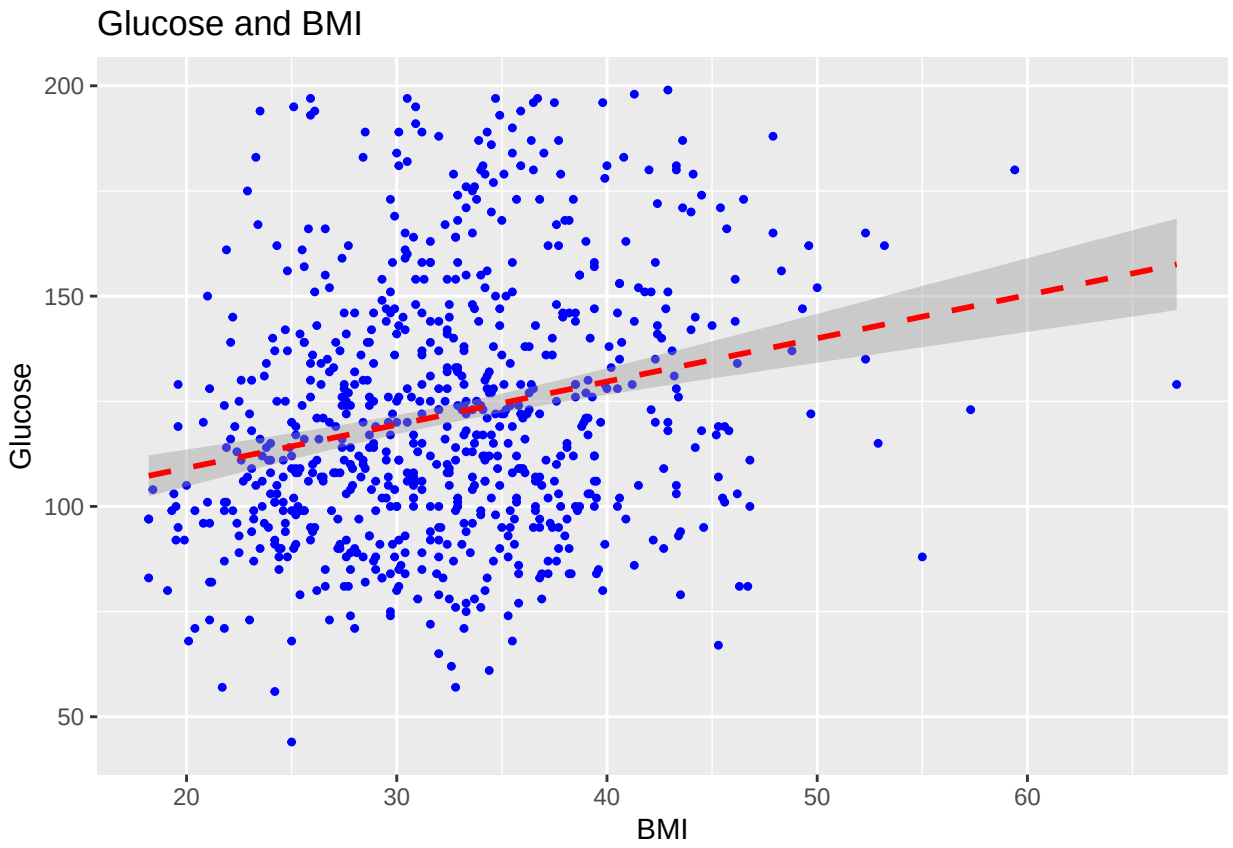
Next we add some color and a smoothing line which helps us see a trend in the data. By default the smoothing line to highlight the trend in the data

```
ggplot(tb, aes(x=mass, y=glucose)) +
  geom_point(pch=20, color='blue', size=1.5) +
  geom_smooth(method='lm', color='red', linetype=2) +
  labs(title="Glucose and BMI", x="BMI", y="Glucose")
```

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

```
## Warning: Removed 16 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 16 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



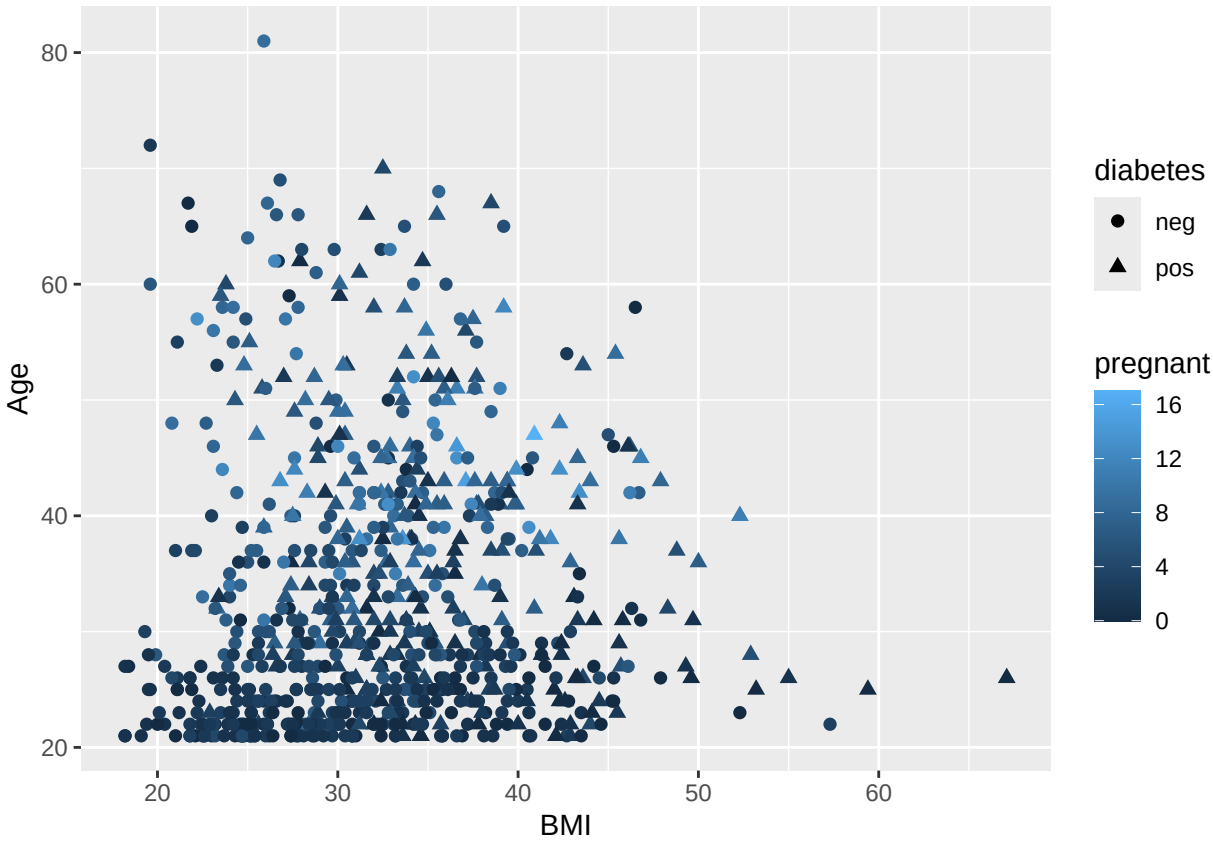
informative graph

```
ggplot(tb,
  aes(x=tb$mass, y=tb$age, shape=diabetes, col=pregnant)) +
  geom_point(size=2) +
  labs(x="BMI",    y="Age")
```

```
## Warning: Use of `tb$mass` is discouraged.
## i Use `mass` instead.
```

```
## Warning: Use of `tb$age` is discouraged.
## i Use `age` instead.
```

```
## Warning: Removed 11 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



facet_grid

- filter out rows with NAs in glucose or insulin
- create 2 new factor columns, glucose_high and insulin_high
- plot

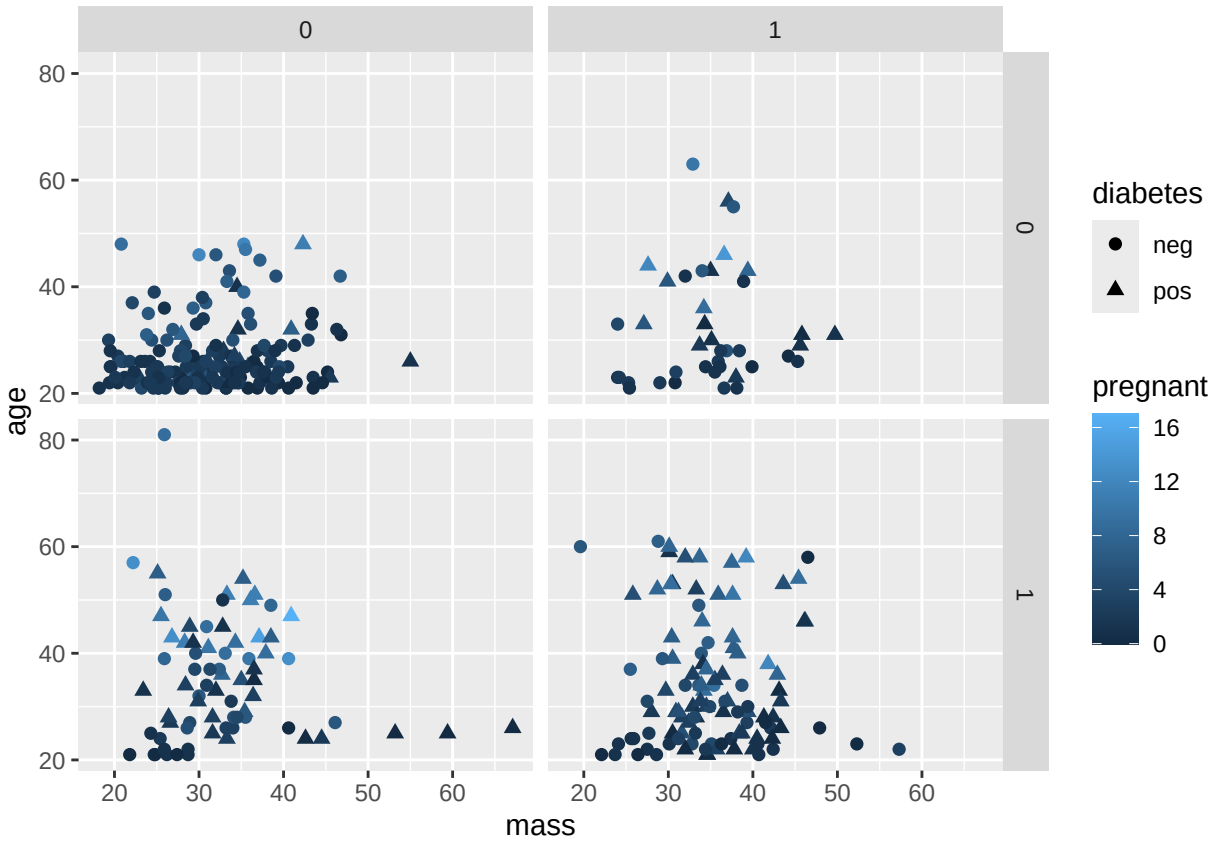
The facet grid for 2 binary variables has 4 windows for all combinations.

```
tb <- filter(tb, !is.na(glucose), !is.na(insulin))

tb <- mutate(tb, glucose_high = factor(ifelse(tb$glucose>mean(tb$glucose), 1, 0)))
tb <- mutate(tb, insulin_high = factor(ifelse(tb$insulin>mean(tb$insulin), 1, 0)))

ggplot(tb,
  aes(x=mass, y=age, shape=diabetes, col=pregnant)) +
  geom_point(size=2) +
 facet_grid(glucose_high~insulin_high)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

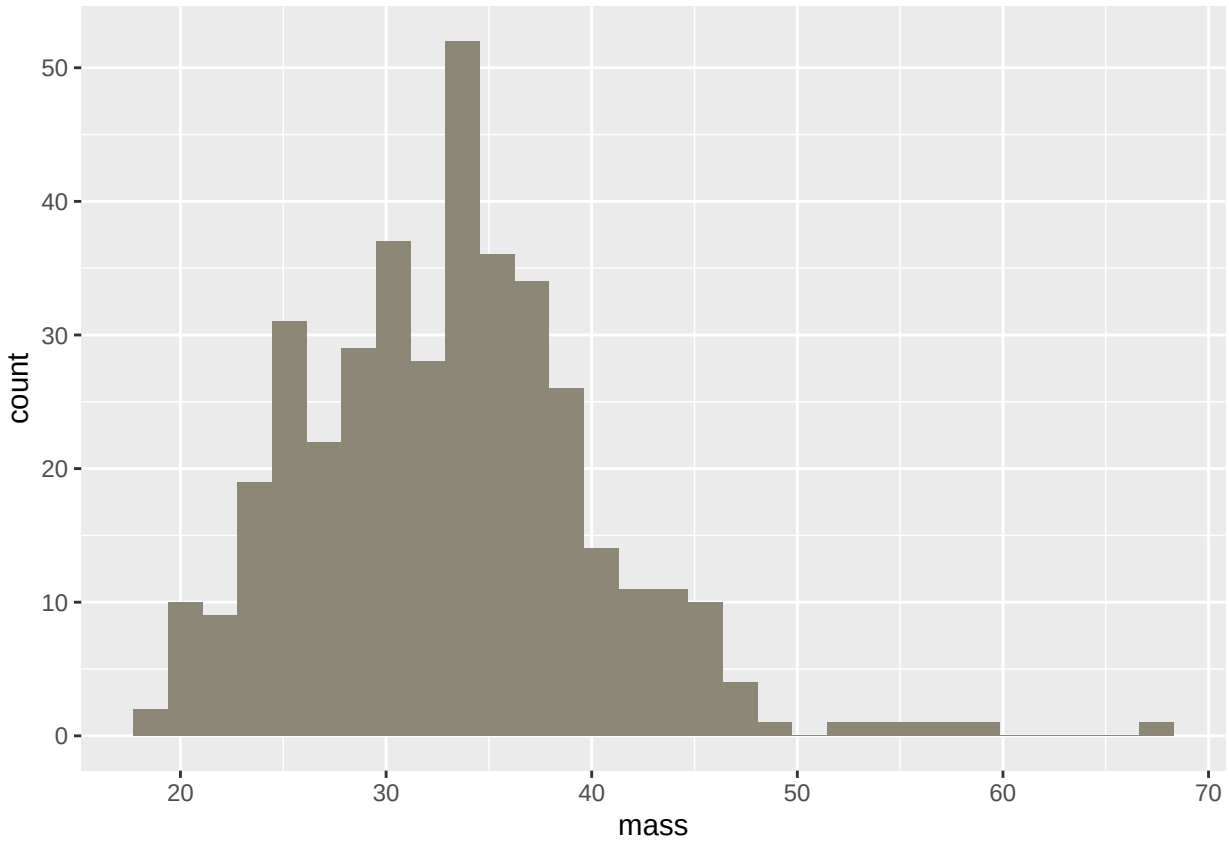


histogram

```
ggplot(tb, aes(x=mass)) +
  geom_histogram(fill="cornsilk4")
```

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

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```

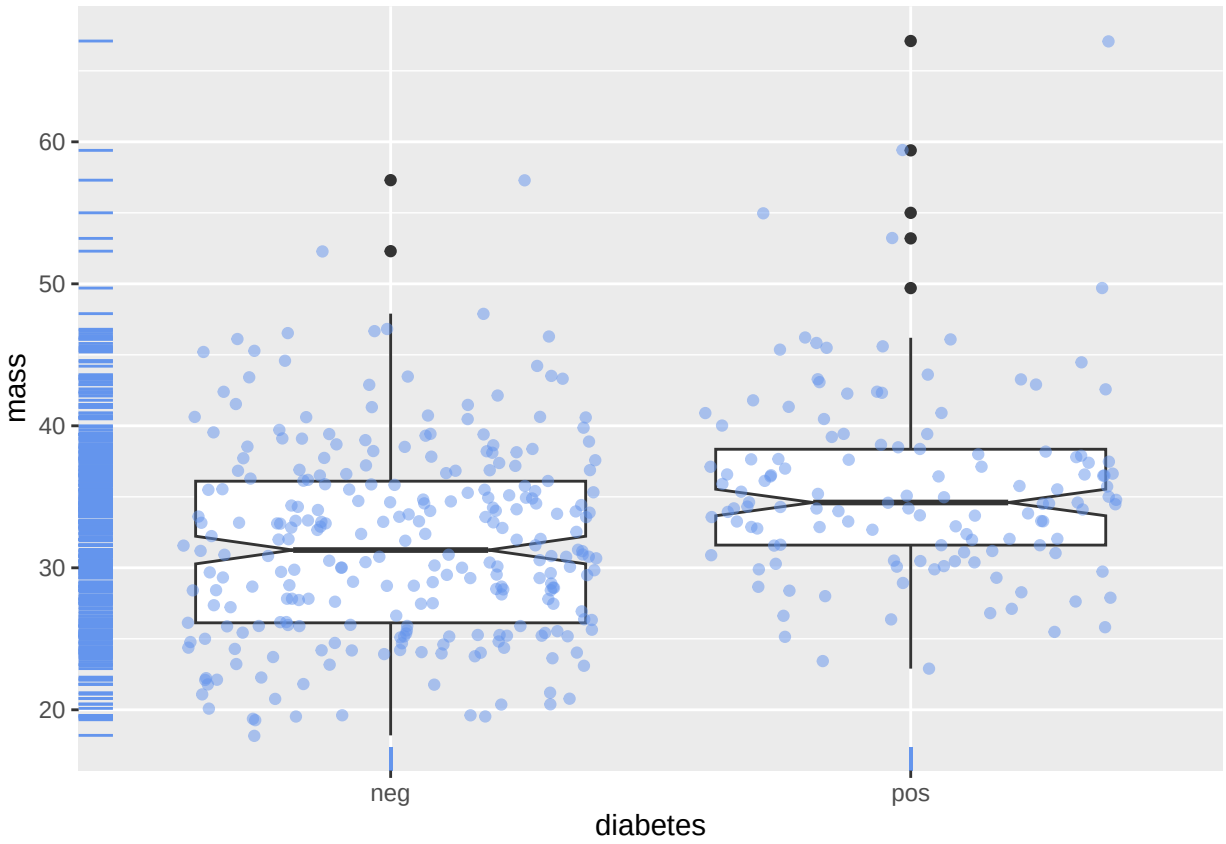


boxplot and rug

```
ggplot(tb, aes(x=diabetes, y=mass)) +
  geom_boxplot(notch=TRUE) +
  geom_point(position="jitter", color="cornflowerblue", alpha=.5) +
  geom_rug(color="cornflowerblue")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_boxplot()`).
```

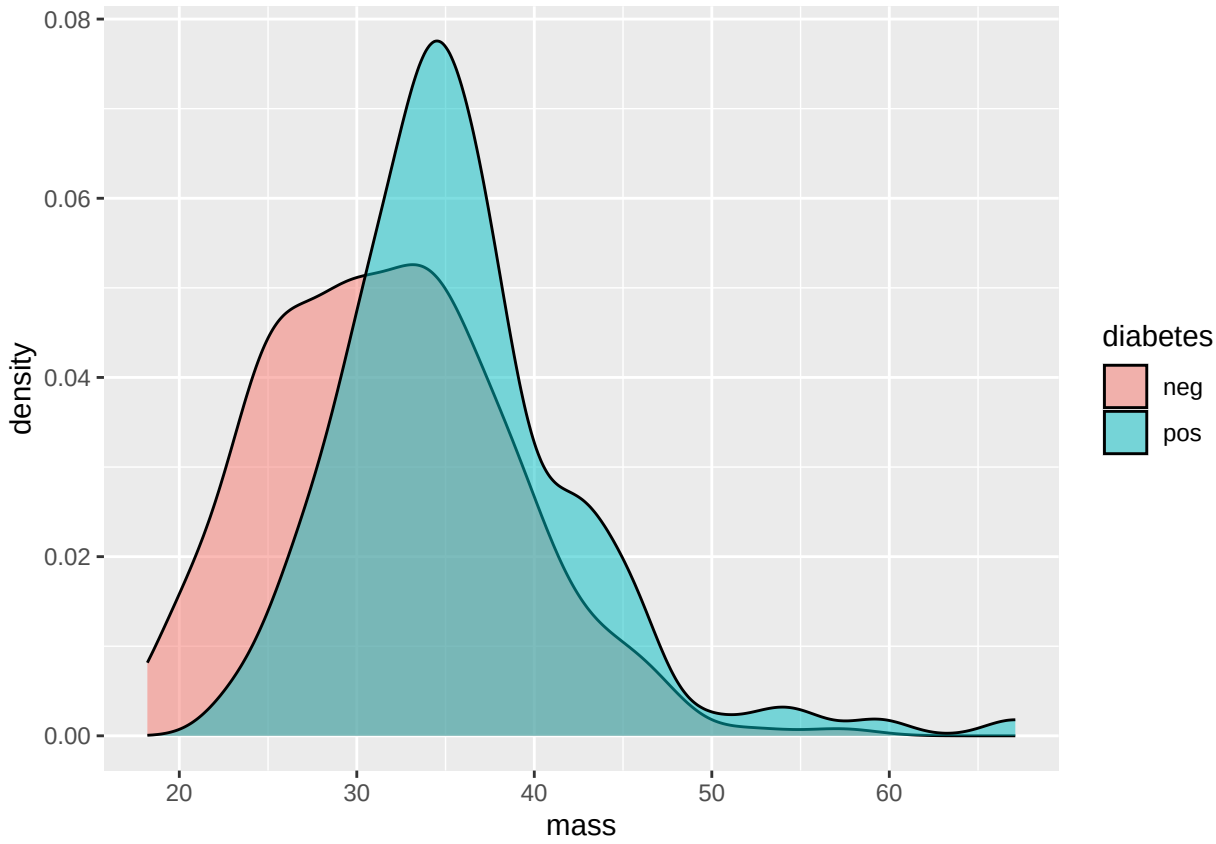
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



density plot

```
ggplot(tb, aes(x=mass, fill=diabetes)) +
  geom_density(alpha=0.5)
```

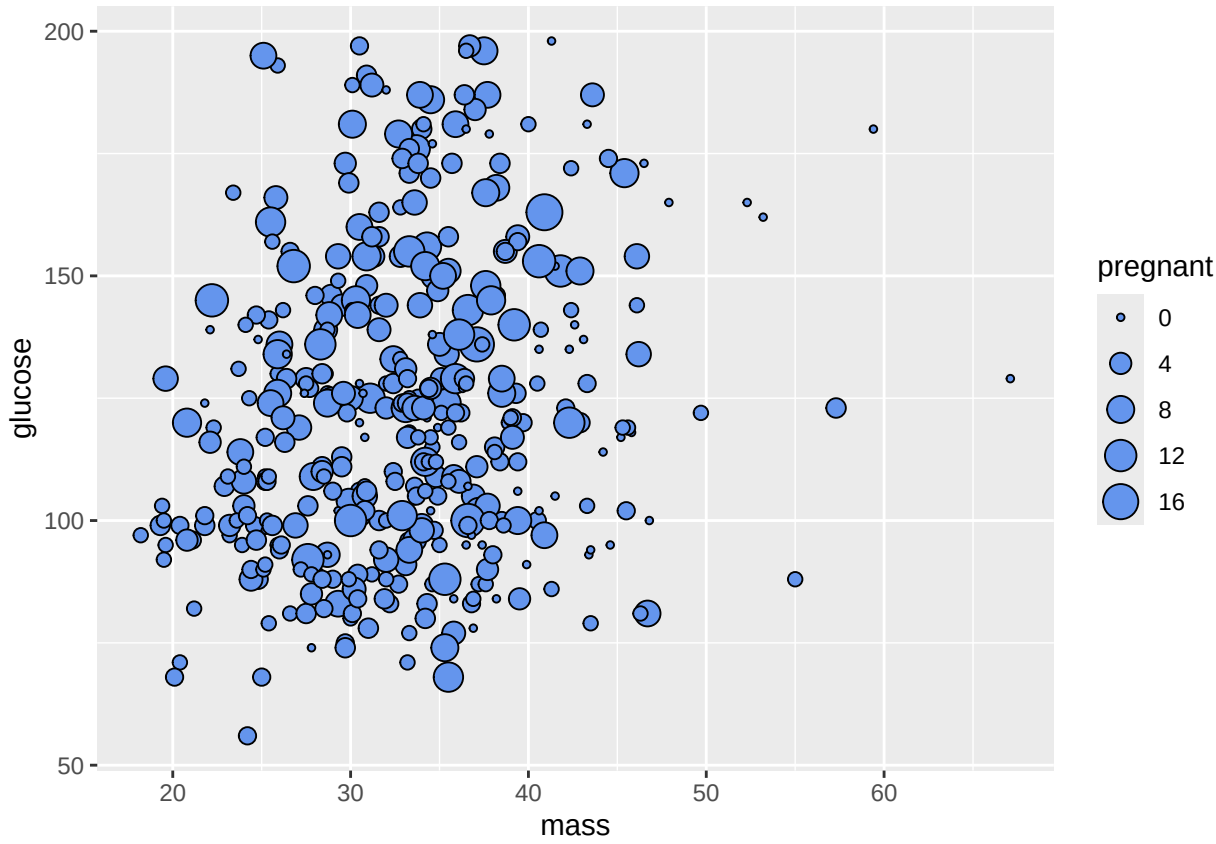
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density()`).
```



bubble chart

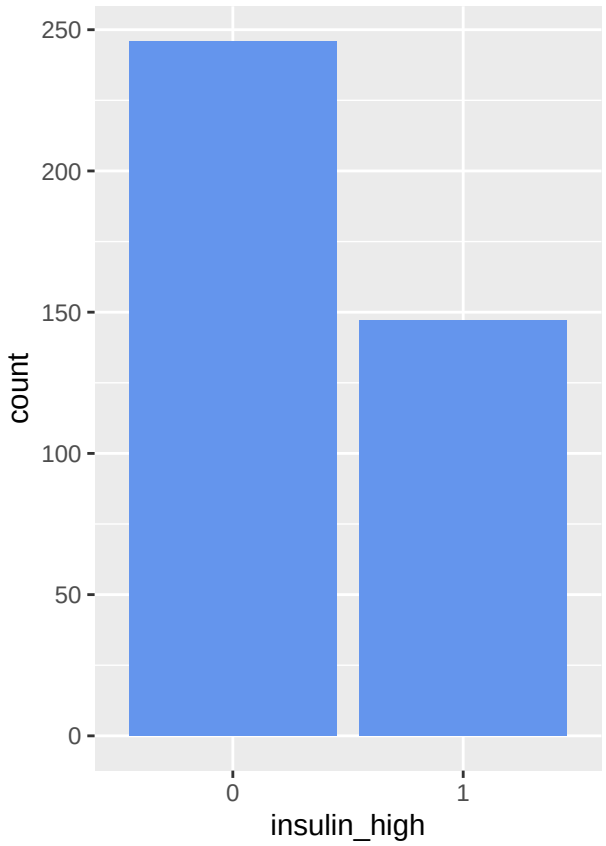
```
ggplot(tb,
       aes(x=mass, y=glucose, size=pregnant)) +
  geom_point(shape=21, fill="cornflowerblue")
```

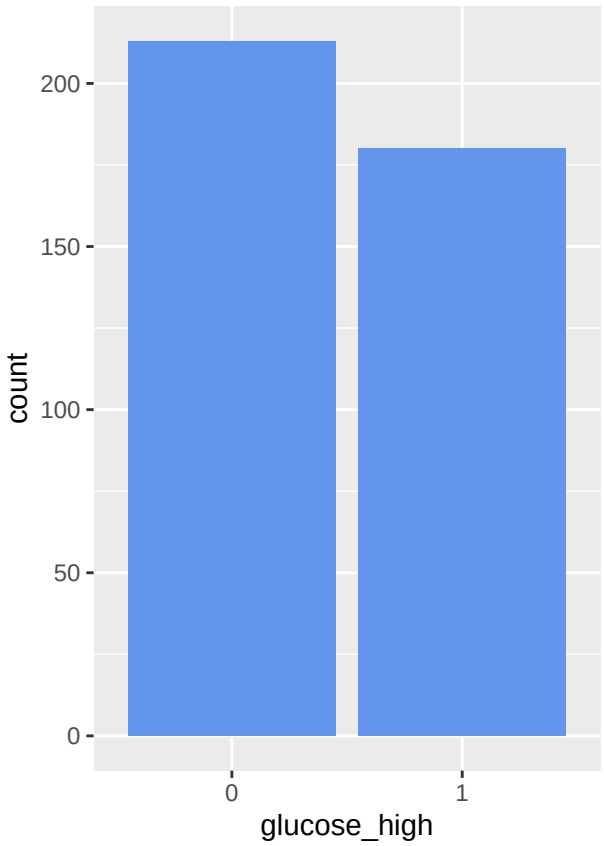
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



grid

```
library(gridExtra)
p1 <- ggplot(tb, aes(x=insulin_high)) + geom_bar(fill="cornflowerblue")
p2 <- ggplot(tb, aes(x=glucose_high)) + geom_bar(fill="cornflowerblue")
grid.arrange(p1, p2, ncol=2)
```


250
200
150
100
50
0
count
0
1
insulin_high


200
150
100
50
0
count
0
1
glucose_high