Introduction to ggplot2

Karen Mazidi

load tidyverse and some data

Loading the diabetes data set from package mlbench.

```
library(tidyverse)
library(mlbench)
data("PimaIndiansDiabetes2")

tb <- tbl_df(PimaIndiansDiabetes2)
```

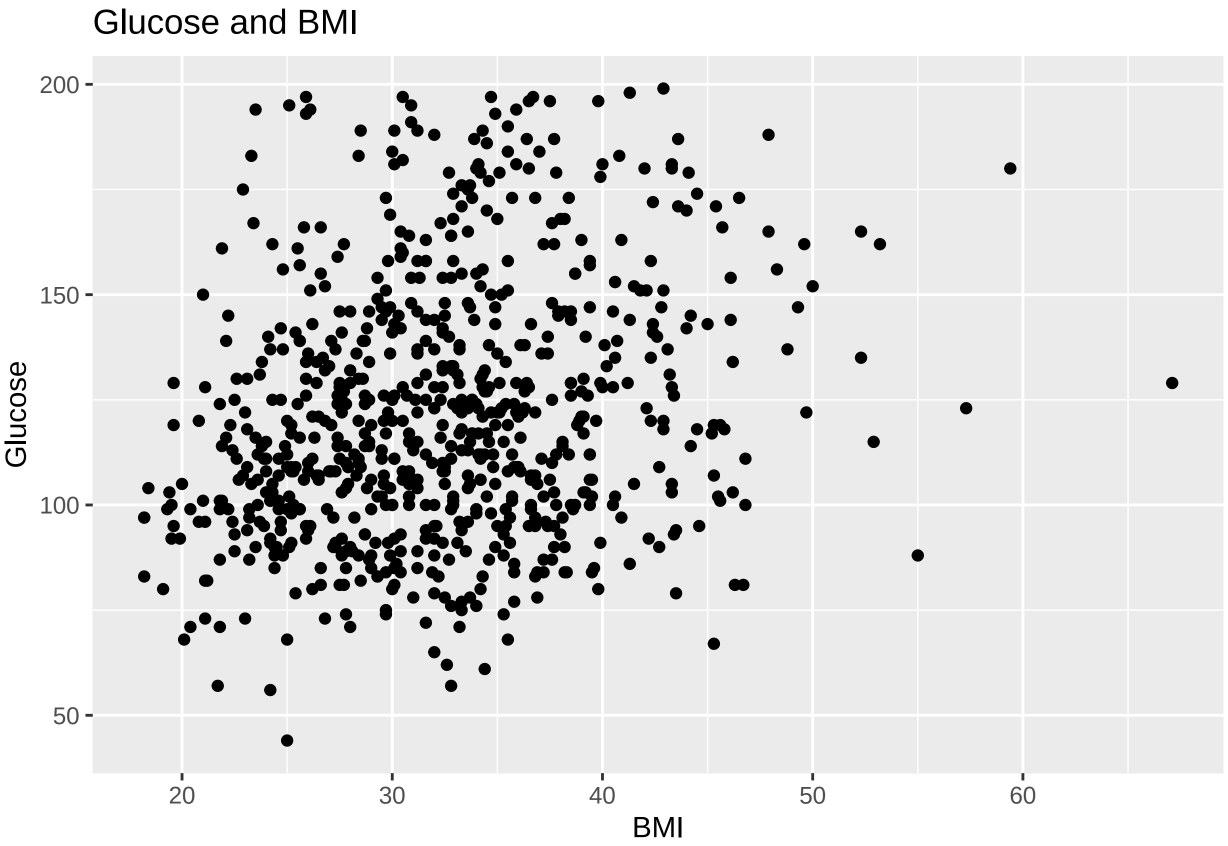
Explore ggplot2

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 (geom_point).
```

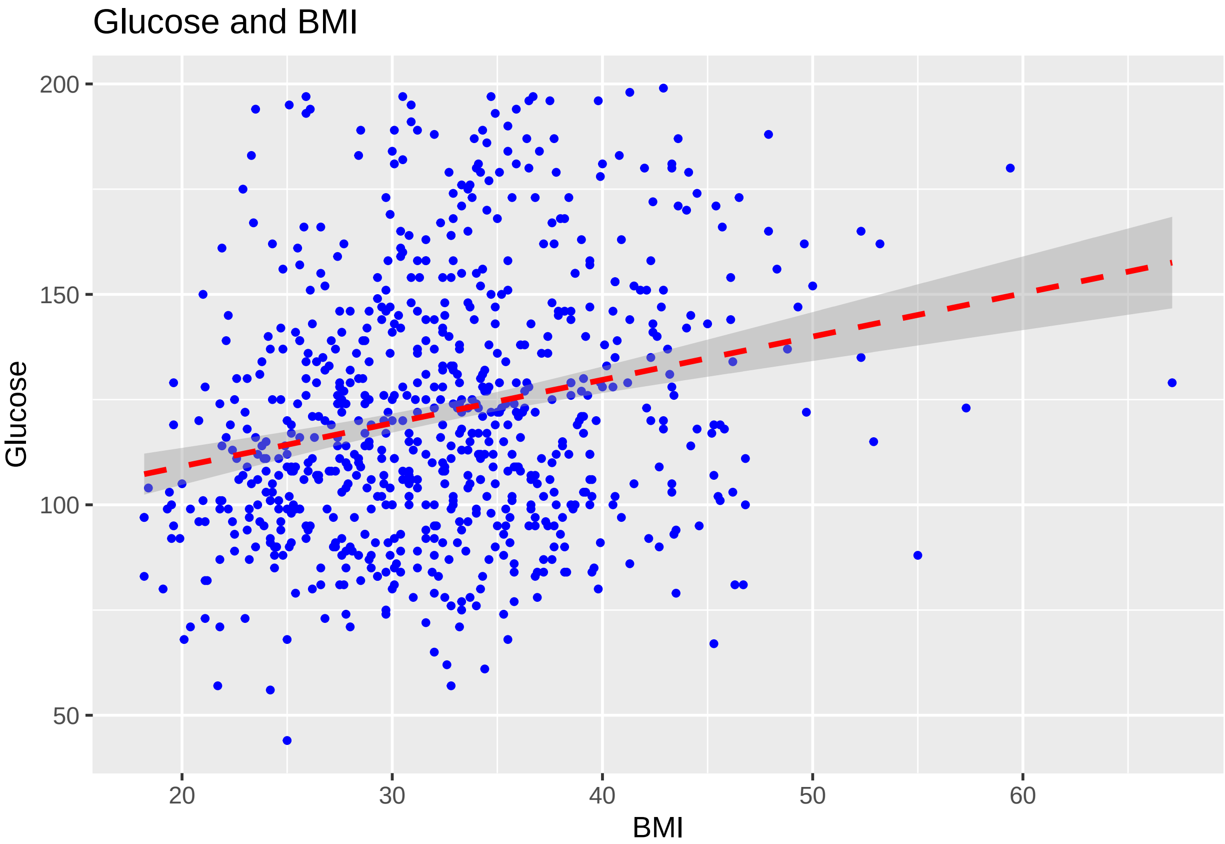


Next we add some color and a smoothing line which helps us see a trend in the data. By default the smoothing line has a shadow around it which specifies the 95

```
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")
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

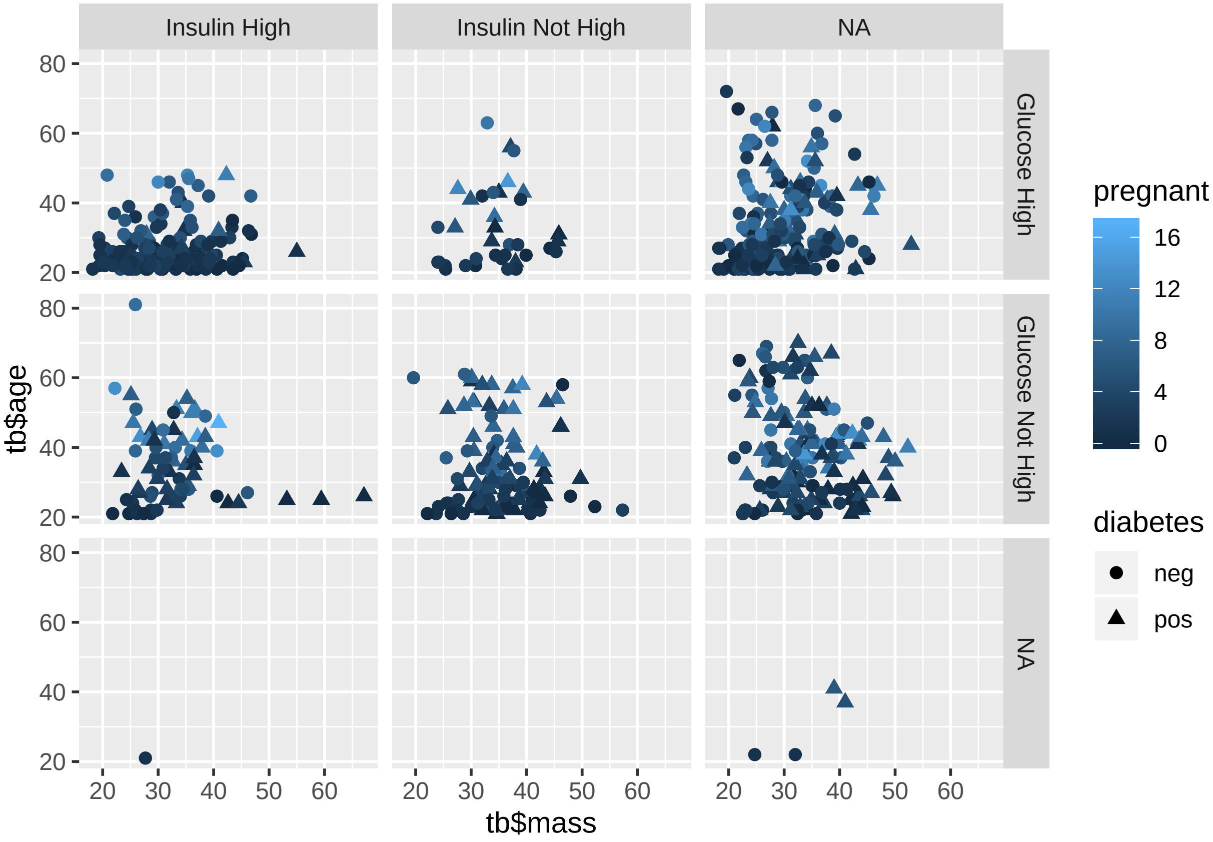


### facet_grid

```
tb$glucose_high <- factor(ifelse(tb$glucose>mean(tb$glucose, na.rm=TRUE), 1, 0),
                          levels=c(0,1), labels=c("Glucose High","Glucose Not High"))
tb$insulin_high <- factor(ifelse(tb$insulin>mean(tb$insulin, na.rm=TRUE), 1, 0),
                          levels=c(0,1), labels=c("Insulin High","Insulin Not High"))

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

```
## Warning: Removed 11 rows containing missing values (geom_point).
```

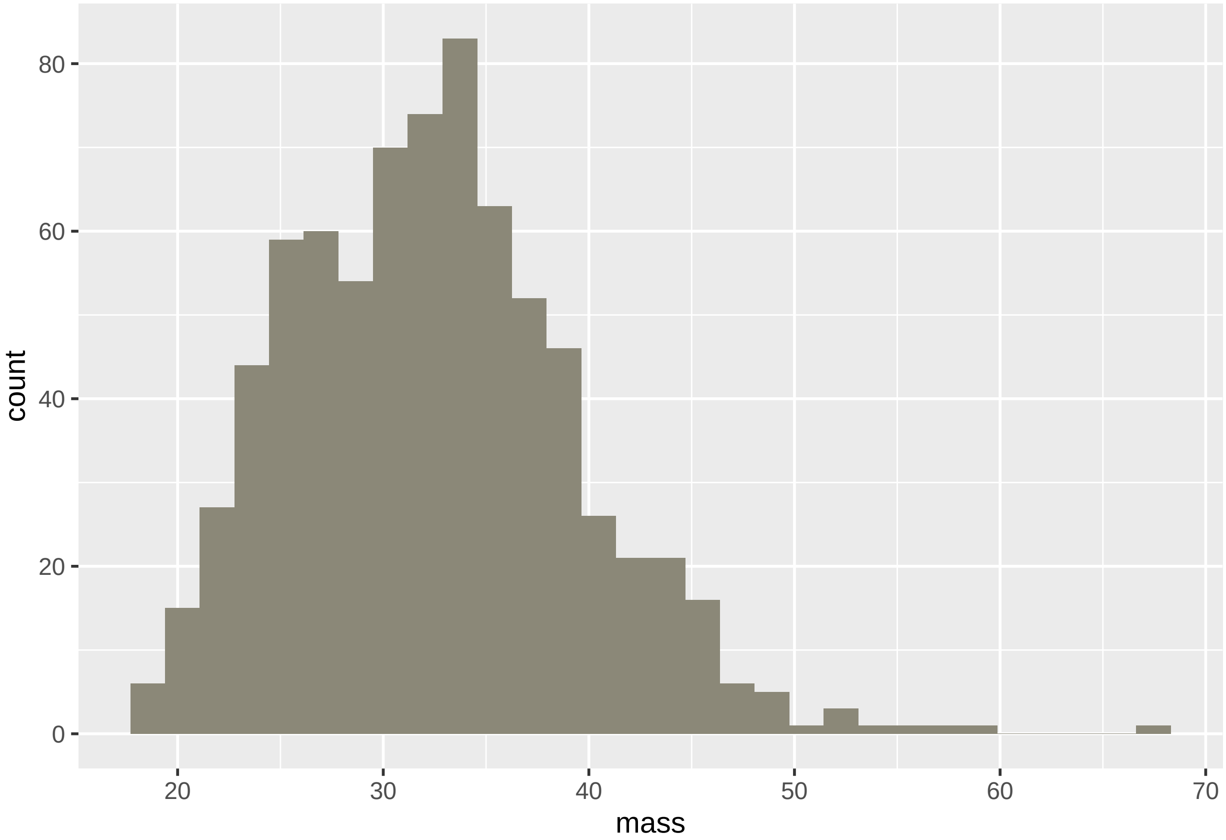


### histogram

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

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

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```

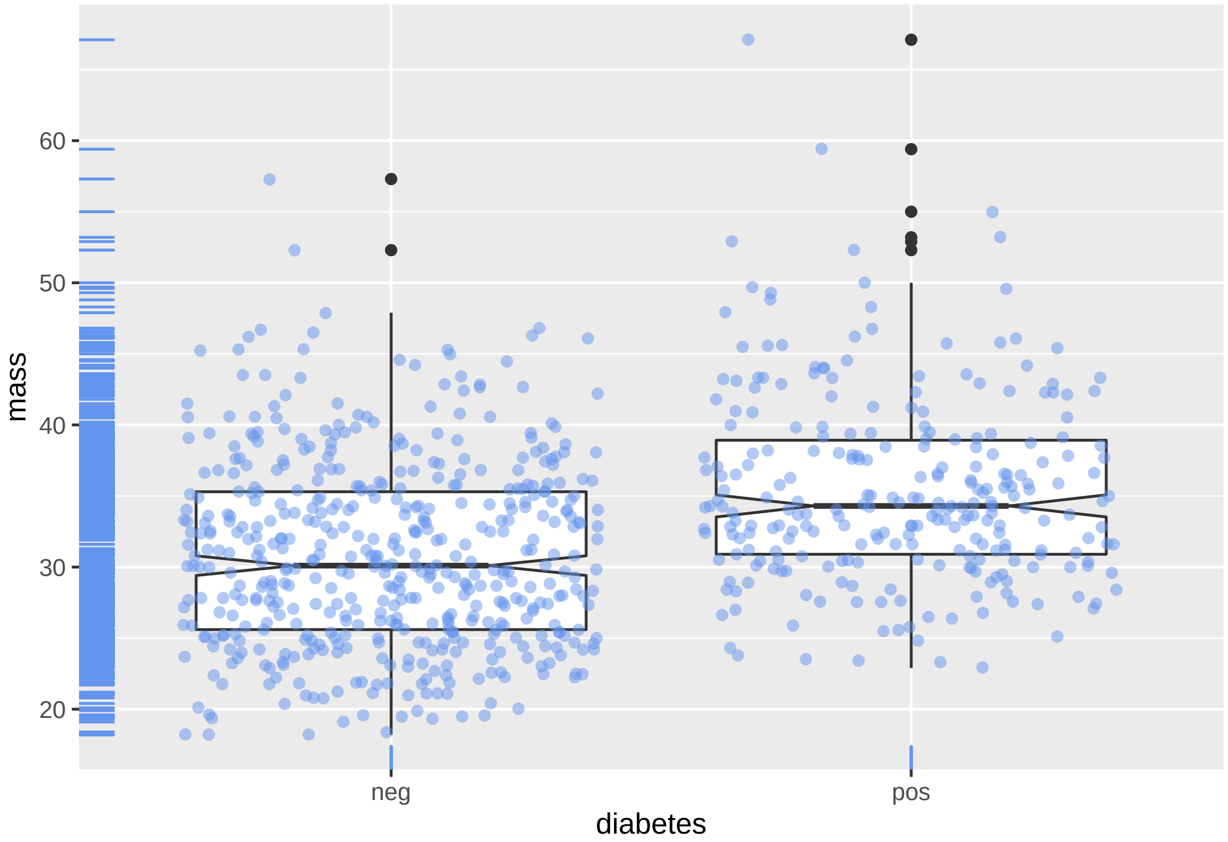


boxplot and rug

```
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 11 rows containing non-finite values (stat_boxplot).
```

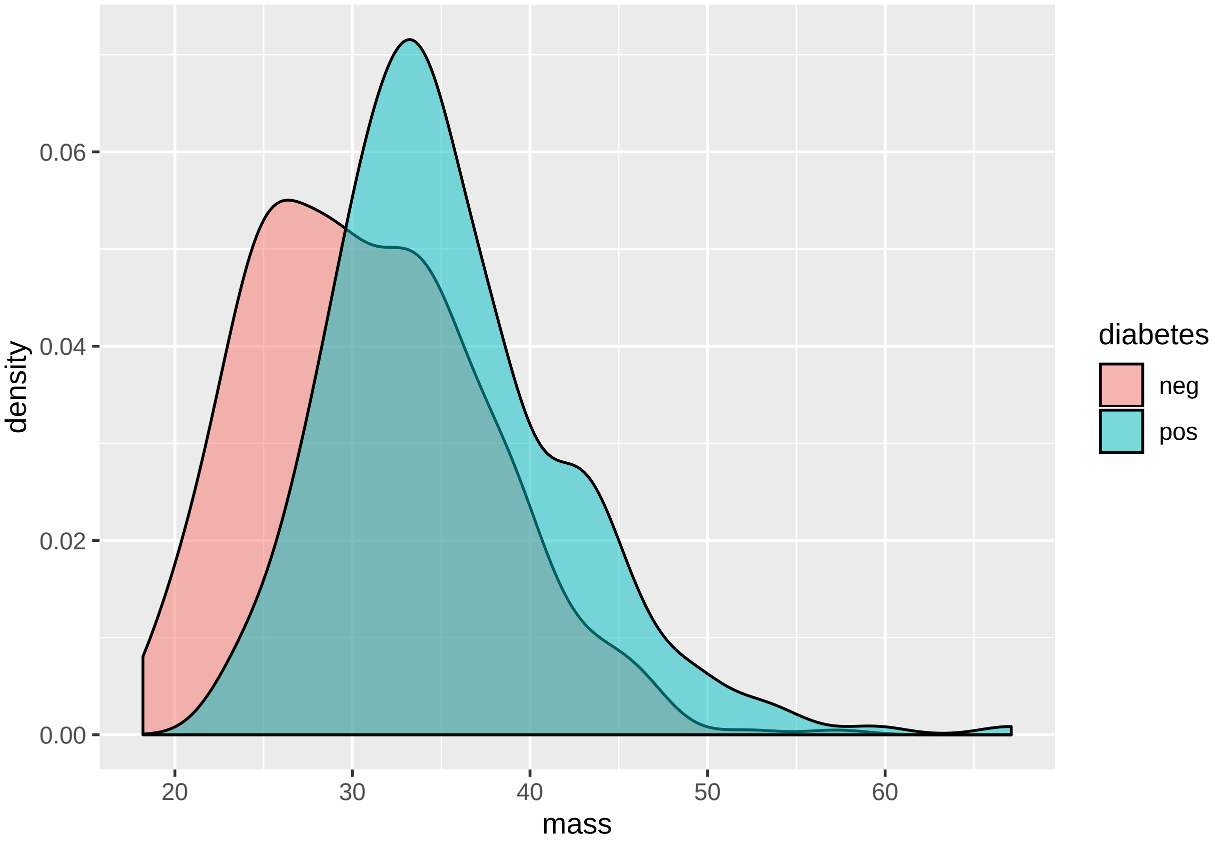
```
## Warning: Removed 11 rows containing missing values (geom_point).
```



density plot

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

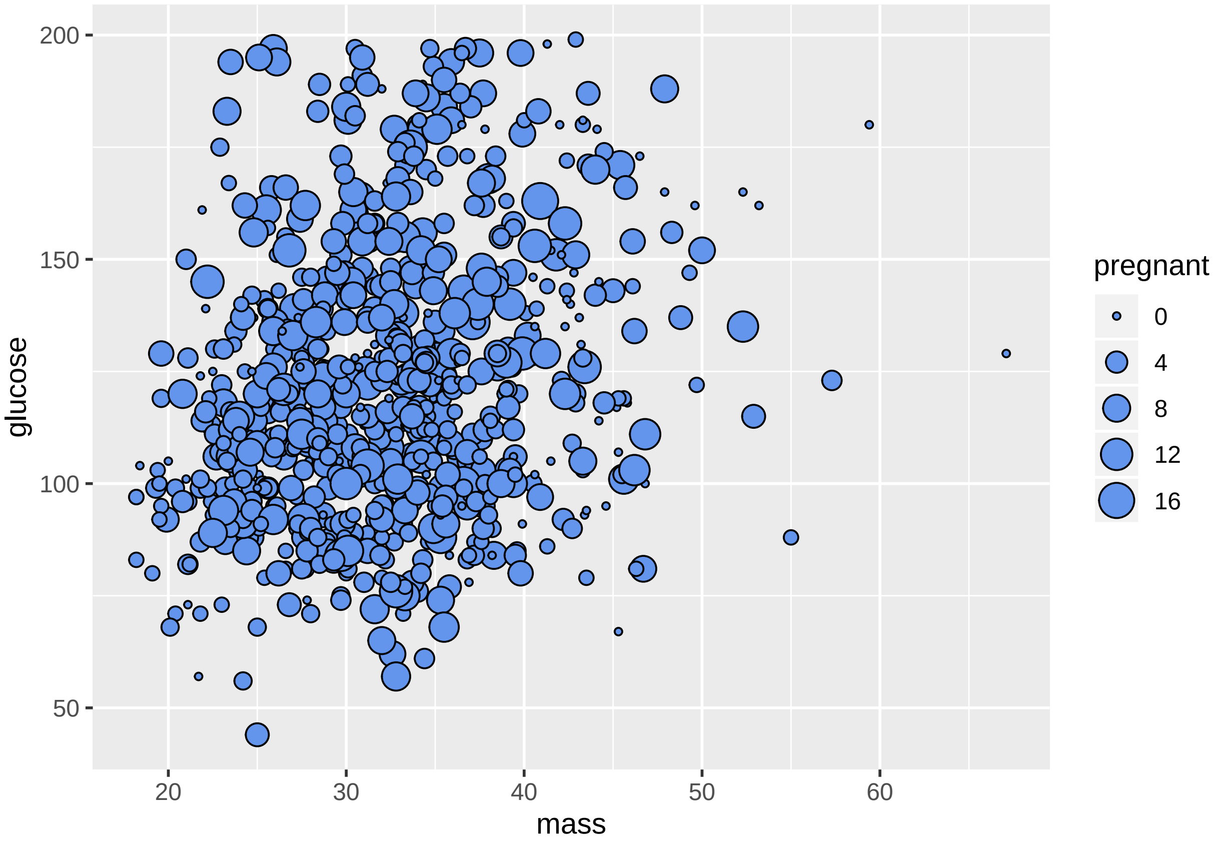
```
## Warning: Removed 11 rows containing non-finite values (stat_density).
```



### bubble chart

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

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



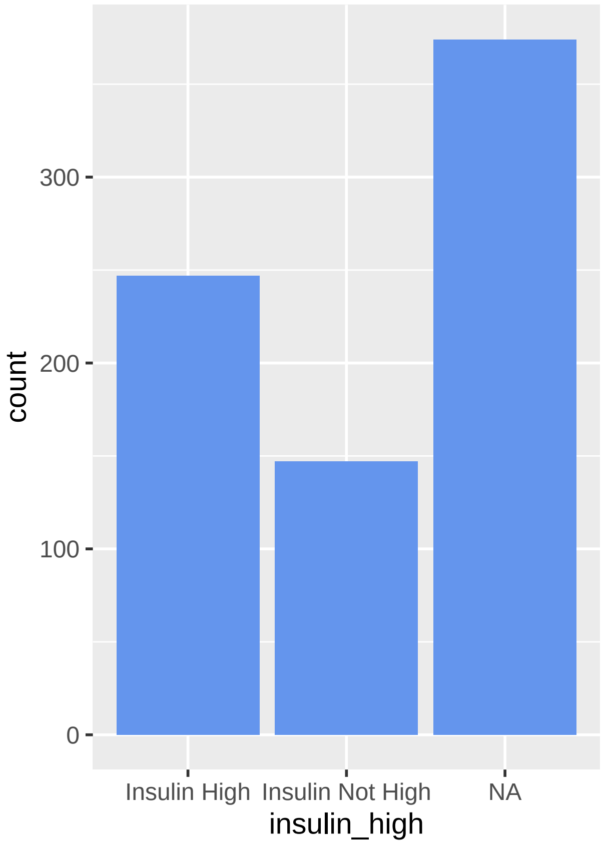
grid

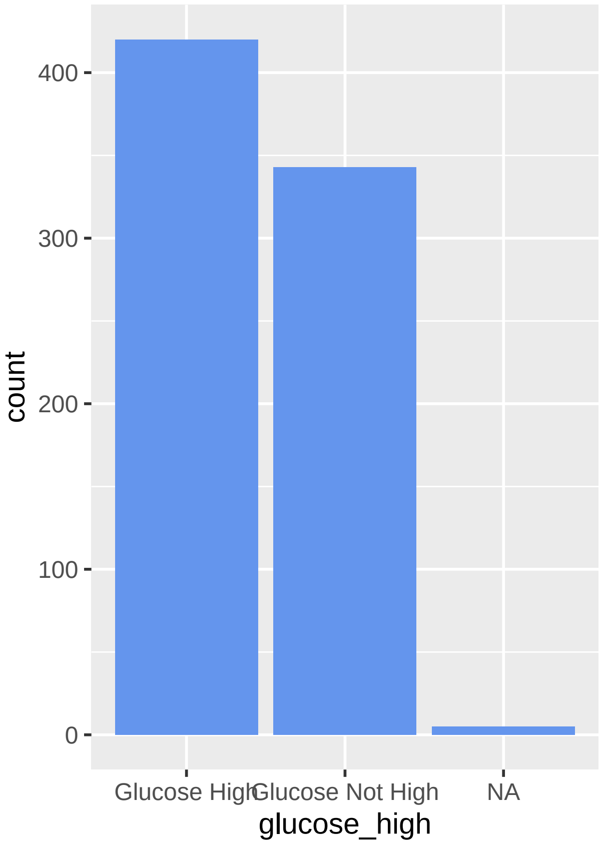
```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```r
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)
```


count
300
200
100
0
Insulin High
Insulin Not High
NA
insulin_high


count
400
300
200
100
0
Glucose High
Glucose Not High
NA
glucose_high