Chapter 1

Graphics with ggplot2 - part 2

1.1 Overplotting

Interger data or data with many categories can cause problems because many points can be at the same position. We can to some extent overcome this by adding some randomness into the plot using **geom_jitter**. We use a the BackPain data for illustration. The data set is large so we use a somewhat reduced random sample and remove the missing data.

```
library(readr)
BackPain<-read_csv("../data/BackPain.csv") # The resulting object is a tibble

# remove missing data in some variables
bp<-BackPain %>% filter(complete.cases(bmi,residence,physical,waistc,height))

set.seed(1001)
bp<-bp[sample(nrow(bp),10000),]
bp

# A tibble: 10,000 x 25
      id residence sex      age wealthQ  physical country  backPain30 agegr maritalS eduS
   <dbl> <chr>     <chr>  <dbl> <chr>    <chr>    <chr>    <chr>      <chr> <chr>    <chr>
 1   679 Rural     Male      56 Q2       high ph~ India    No         50-59 Married~ No p~
 2  8863 Rural     Female    56 Q3       high ph~ China    No         50-59 Married~ No p~
 3 33223 Urban     Male      58 Q4       high ph~ Russian~ Yes        50-59 Married~ Comp~
 4 30317 Urban     Female    73 Q2       high ph~ China    Yes        70-79 Married~ No p~
 5 15269 Urban     Male      53 Q5 rich~ mod phy~ China    No         50-59 Married~ Comp~
 6 30488 Urban     Female    69 Q5 rich~ mod phy~ Russian~ Yes        60-69 Div/Wid~ Comp~
 7 20081 Rural     Male      71 Q2       low phy~ Russian~ No         70-79 Married~ Comp~
 8  5997 Urban     Female    60 Q5 rich~ mod phy~ China    No         60-69 Div/Wid~ Comp~
 9 31587 Urban     Female    72 Q3       mod phy~ China    Yes        70-79 Married~ Comp~
10 13102 Rural     Male      64 Q4       low phy~ Mexico   No         60-69 Married~ Comp~
# ... with 9,990 more rows, and 14 more variables: workS <chr>, bmi <dbl>, bmi4 <chr>,
#   waistc <dbl>, smoke <chr>, alcohol <chr>, arthritis <chr>, angina <chr>,
#   depression <chr>, asthma <chr>, diabetes <chr>, comorb <dbl>, disability <dbl>,
#   height <dbl>

p<-ggplot(bp,aes(y=disability,x=age))
p1<-p+geom_point(size=0.7)
# Not informative with integer or lattice data. There are points at the same position.
p2<-p+geom_jitter(height=0.5,width=0.5,size=1,shape=21)  # Better
```

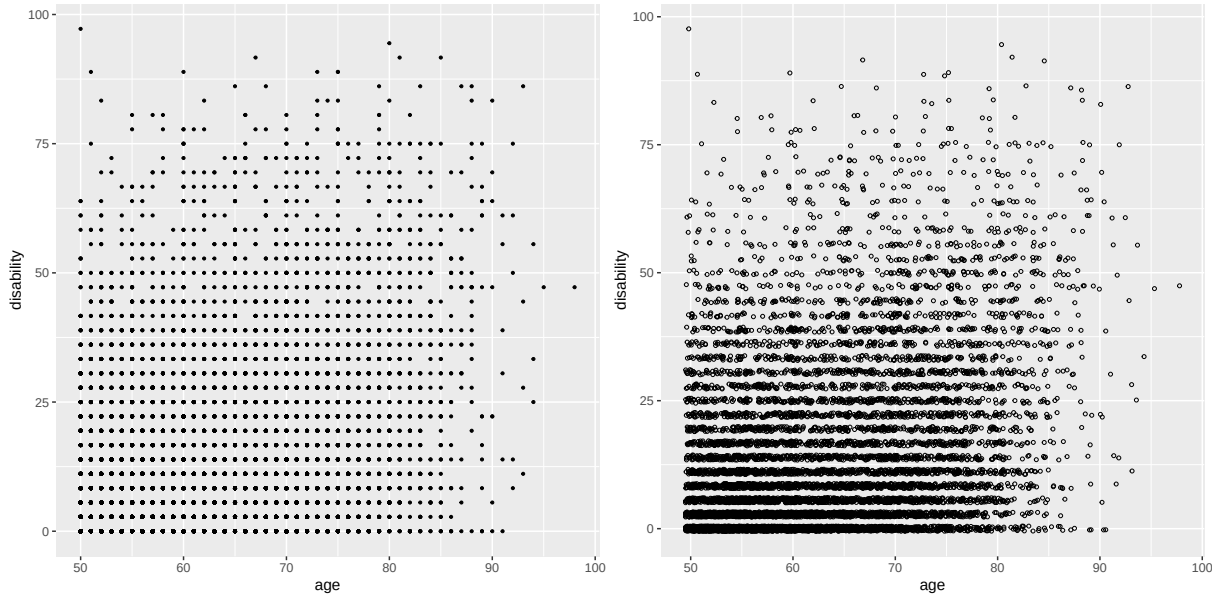

Figure 1.1: Scatterplots; normal (p1) and using geom_gitter (p2)

1.2 Multiple plots

It is often necessary to illustrate subsets in the same plot. They can be separated using different lines, symbols and color.

Note:

Don't forget that you can always use the R-studio menu Plots/Zoom or Plots/Export to get a better view of the plot.

1.2.1 Using different data sets in the same plot

Let us now import the data file norsjo86.

agegrp:	Age group	(30, 40, 50 ,60 years)
health:	Health status	(0=good, 1=not quite good/bad)
sex:	Sex	(1=man, 2=woman)
height:	Body height	(cm)
weight:	Body weight	(kg)
sbp:	Systolic blood pressure	
dbp:	Diastolic blood pressure	
cholesterol:	Cholesterol	
smoker:	Smoking status	(0=non-smoker, 1=smoker)
bmi:	Body mass index	(kg/m^2)

```
library(haven)
norsjo86 <- read_sav("../data/norsjo86.sav")

norsjo86<- norsjo86 %>% filter(complete.cases(sbp,bmi,cholesterol))
```

```
Error:  Problem with `filter()` input `..1`.
x object 'cholesterol' not found
i Input `..1` is `complete.cases(sbp, bmi, cholesterol)`.
```

```
names(norsjo)[names(norsjo) == "kolester"] <- "cholesterol" # rename
norsjo86
```

```
# A tibble: 260 x 10
         agegrp          health      sex height weight   sbp   dbp kolester    smoker   bmi
      <dbl+lbl>      <dbl+lbl> <dbl+lb>  <dbl>  <dbl> <dbl> <dbl>    <dbl> <dbl+lbl> <dbl>
 1 60 [60 year~ 0 [good]       2 [Woma~    157     61   110    70      6.7 0 [Non-s~  24.7
 2 60 [60 year~ 1 [not quite ~ 2 [Woma~    157     97   150   100      6.6 0 [Non-s~  39.4
 3 60 [60 year~ 0 [good]       1 [Man]     170     74   136    96      8.2 0 [Non-s~  25.6
 4 60 [60 year~ 0 [good]       2 [Woma~    163     66   156    76      7.5 0 [Non-s~  24.8
 5 60 [60 year~ 0 [good]       2 [Woma~    166     66   110    70     10.2 0 [Non-s~  24.0
 6 60 [60 year~ 0 [good]       2 [Woma~    168     61   130    78      7.3 0 [Non-s~  21.6
 7 60 [60 year~ 1 [not quite ~ 2 [Woma~    159     67   122    74      5.2 0 [Non-s~  26.5
 8 60 [60 year~ 0 [good]       1 [Man]     172     62   142    88      7.4 0 [Non-s~  21.0
 9 60 [60 year~ 1 [not quite ~ 2 [Woma~    153     68   150   100      7.2 0 [Non-s~  29.0
10 60 [60 year~ 1 [not quite ~ 1 [Man]     179     87   133    85      7.8 0 [Non-s~  27.2
# ... with 250 more rows
```

```
gr1<-norsjo86 %>% filter(sex==1 & smoker==0)
gr2<-norsjo86 %>% filter(sex==2 & smoker==1)

p1<-ggplot(gr1,aes(y=sbp,x=bmi))+
   geom_point(size=1.5,color="blue")+
   geom_smooth(method=lm,se=F,color="blue")+
   labs(title="Non-smoking men (p1)")

p2<-p1+geom_point(data=gr2,size=1.5,color="red")+
   geom_smooth(data=gr2,method=lm,se=F,color="red")+
   labs(title="Smoking women (red) and non-smoking men ( blue) (p2)")
```

In the second plot new data were added in the geoms.

1.2.2 Using aestetics (aes argument) to make multiple plots

The different parameters in aestetics can be found at
https://cran.r-project.org/web/packages/ggplot2/vignettes/ggplot2-specs.html

Histogram

```
p<-ggplot(bp,aes(x=bmi))
p1<-p+geom_histogram(binwidth=1,colour="black",fill="red",alpha=0.4) # alpha=1 is full color

p2<-p+geom_histogram(aes(fill=sex),binwidth=1,colour="black",alpha=1)
```

Boxplots

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## Warning: Removed 3 rows containing missing values (geom_point).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 3 rows containing missing values (geom_point).
```

Figure 1.2: Scatterplot with regression line, one and two subgroups

```
p<-ggplot(bp,aes(x=country,y=bmi))
p1<-p+geom_boxplot(colour="blue",size=0.5)+ # size=thickness of box lines
    labs(title="p1")
p2<-p+geom_boxplot(aes(fill=sex)) +
   labs(title="p2")
p3<-p+geom_boxplot(aes(fill=agegr),alpha=0.8)+
   labs(title="p3")
p4<-p+geom_boxplot(aes(color=agegr))+
   labs(title="p4")
```

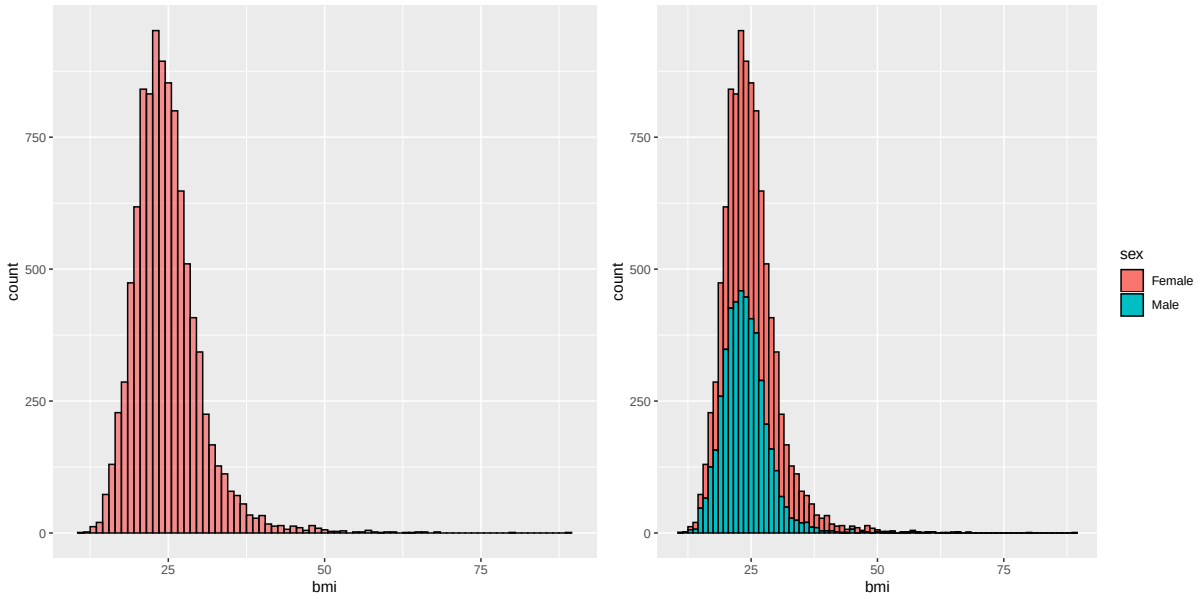

Figure 1.3: Histograms; total (p1) and with aes(fill=sex) (p2)

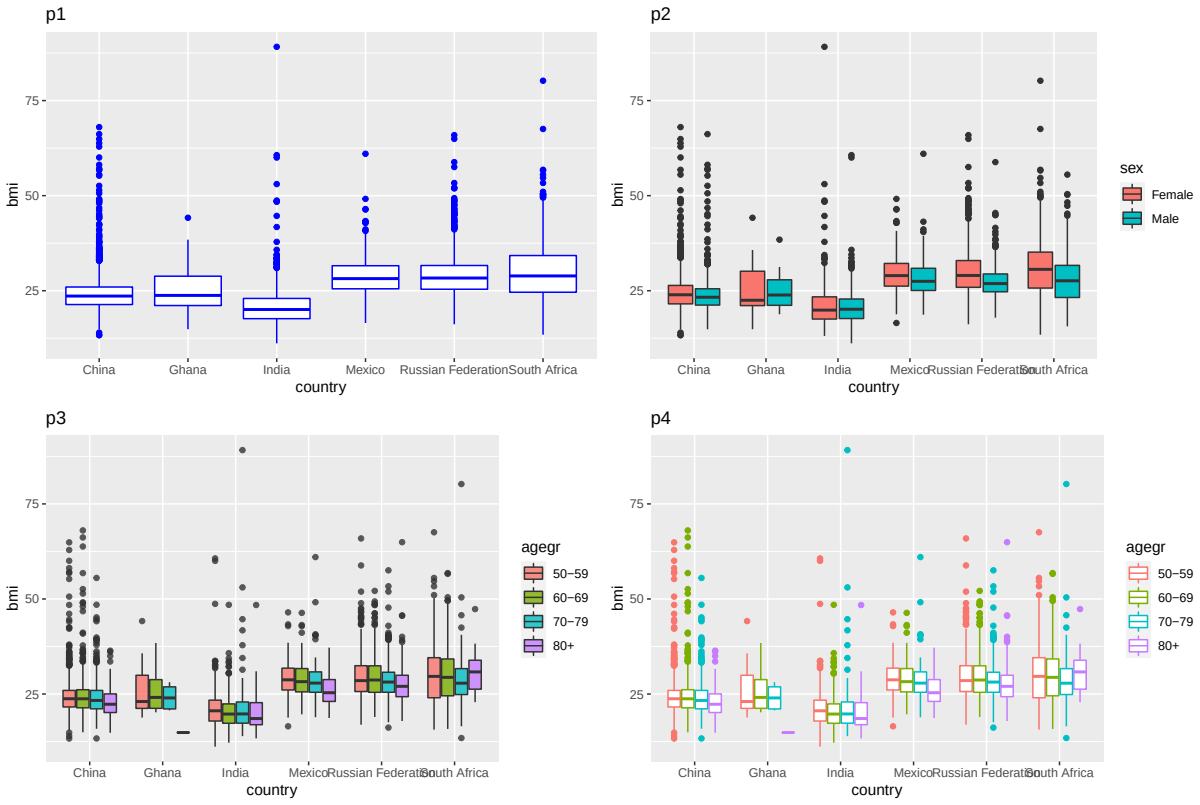

Figure 1.4: Boxplots; examples of aestetic mappings on agegroup and sex

Scatterplots

```
norsjo86<-norsjo86 %>% mutate(Sex=as_factor(sex))

p<-ggplot(norsjo86,aes(y=sbp,x=bmi,shape=Sex))
#p<-ggplot(norsjo86,aes(y=sbp,x=bmi,shape=as_factor(sex)))
#this alternative works but does not look as good
p1<-p+geom_point(size=1.5,aes(color=Sex))+
    labs(title="p1")
p2<-p+geom_point(aes(size=cholesterol/2))+
 labs(title="p2")
p3<-p+geom_point(shape=21,colour="black",fill="red",aes(size=cholesterol/2,stroke=bmi/10))+
     labs(title="p3")
# we can add regression lines
p4<-p1+geom_smooth(method=lm,color="black",linetype=2,se=F,aes(color=Sex))+
 labs(title="p4")
```

In figure p3 `aes stroke` is using bmi which is also the x-variable. This is not how it should be used but the purpose here is to demonstrate how it works.

```
## Warning:  Removed 5 rows containing missing values (geom_point).
## Error in FUN(X[[i]], ...):  object 'cholesterol' not found
```

Figure 1.5: Scatterplots; examples of aestetics

Trend plots

A trend plot is a plot where values on the x-axis is ordinal, e.g. time in years and there is only one y-value per x-value.

```
norsjo86<-norsjo86 %>% mutate(sex=factor(sex,labels=c("men","women")),
                              age=as.double(agegrp))
nor.sum<-norsjo86 %>% group_by(age,sex) %>%
   summarise(mean.sbp=mean(sbp),min.sbp=min(sbp),max.sbp=max(sbp))
nor.sum

# A tibble: 8 x 5
# Groups:   age [4]
    age sex   mean.sbp min.sbp max.sbp
  <dbl> <fct>    <dbl>   <dbl>   <dbl>
1    30 men       119        90     148
2    30 women     110.       97     142
3    40 men       121.       98     152
4    40 women      NA        NA      NA
5    50 men       127.      102     220
6    50 women     132.       93     176
7    60 men        NA        NA      NA
8    60 women      NA        NA      NA

p1<-ggplot(nor.sum,aes(y=mean.sbp,x=age,shape=sex))+
   geom_line(aes(linetype=sex),size=1)+
   geom_point(size=2)+
   labs(y="Mean systolic blood pressure")+
   labs(title="p1")
```

```
p2<-p1+geom_point(aes(y=max.sbp),color="black",size=2)+
   geom_line(aes(y=max.sbp,linetype=sex),size=1)+
   geom_point(aes(y=min.sbp),color="black",size=2)+
    geom_line(aes(y=min.sbp,linetype=sex),size=1)+
   labs(y="Systolic blood pressure",title="Mean, maximum and minimum systolic blood pressure",subtitl

p3<-p1+geom_point(color="black")+
   geom_text(aes(y=min.sbp,color=sex),label="Min",size=3)+
   geom_label(aes(y=max.sbp,color=sex),label="Max",size=3)+
   labs(title="p3")
# label can also be a vector with length=nrow(data.frame)
```

```
## Warning:  Removed 2 row(s) containing missing values (geom_path).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 2 row(s) containing missing values (geom_path).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 2 row(s) containing missing values (geom_path).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 2 row(s) containing missing values (geom_path).
## Warning:  Removed 2 row(s) containing missing values (geom_path).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 3 rows containing missing values (geom_point).
## Warning:  Removed 3 rows containing missing values (geom_text).
## Warning:  Removed 3 rows containing missing values (geom_label).
```

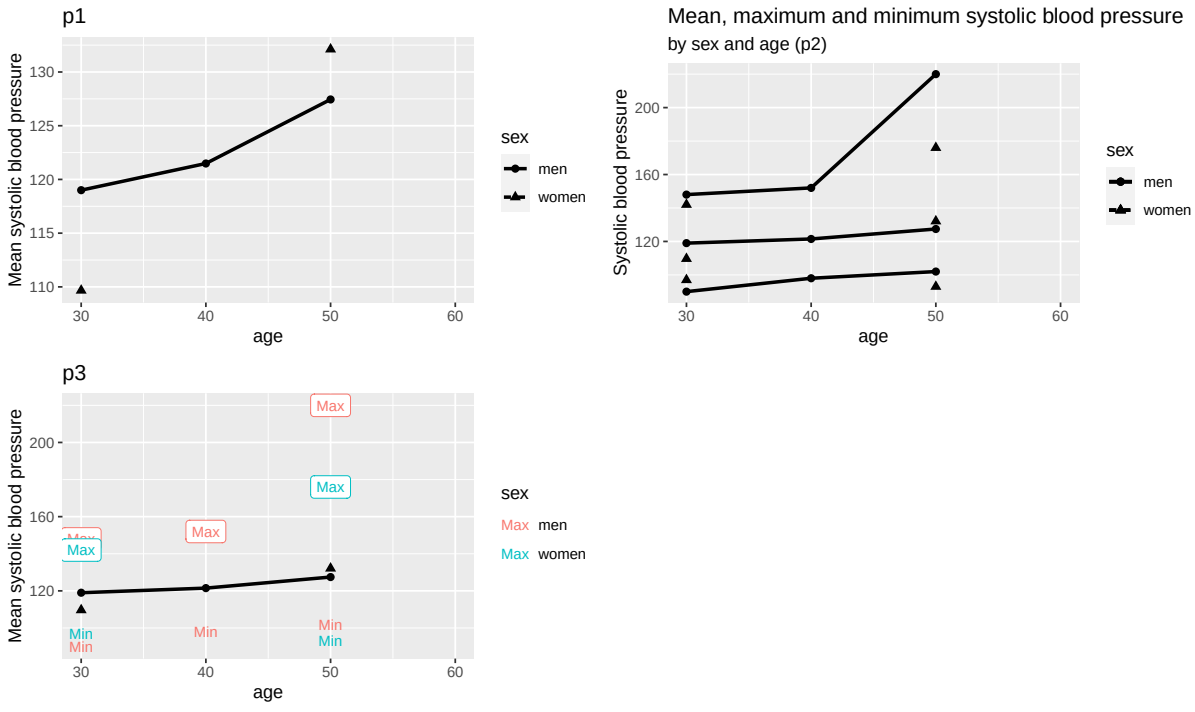

Figure 1.6: Trend plots including summary measures

Interaction

We can also use interaction of two factors in ggplot and in aestetics.

```
p<-ggplot(bp,aes(x=interaction(sex,residence),y=bmi))
p1<-p+geom_boxplot(colour="black",size=0.5)+
    labs(title="p1")
p2<-p+geom_boxplot(size=0.5,aes(fill=sex))+
    labs(title="p2")
p3<-p+geom_boxplot(colour="blue",size=0.5,aes(fill=interaction(sex,residence)))+
    labs(title="p3")
p<-ggplot(bp,aes(x=agegr,y=bmi))
p4<-p+geom_boxplot(colour="blue",size=0.5,aes(fill=interaction(sex,residence)))+
    labs(title="p4")
```

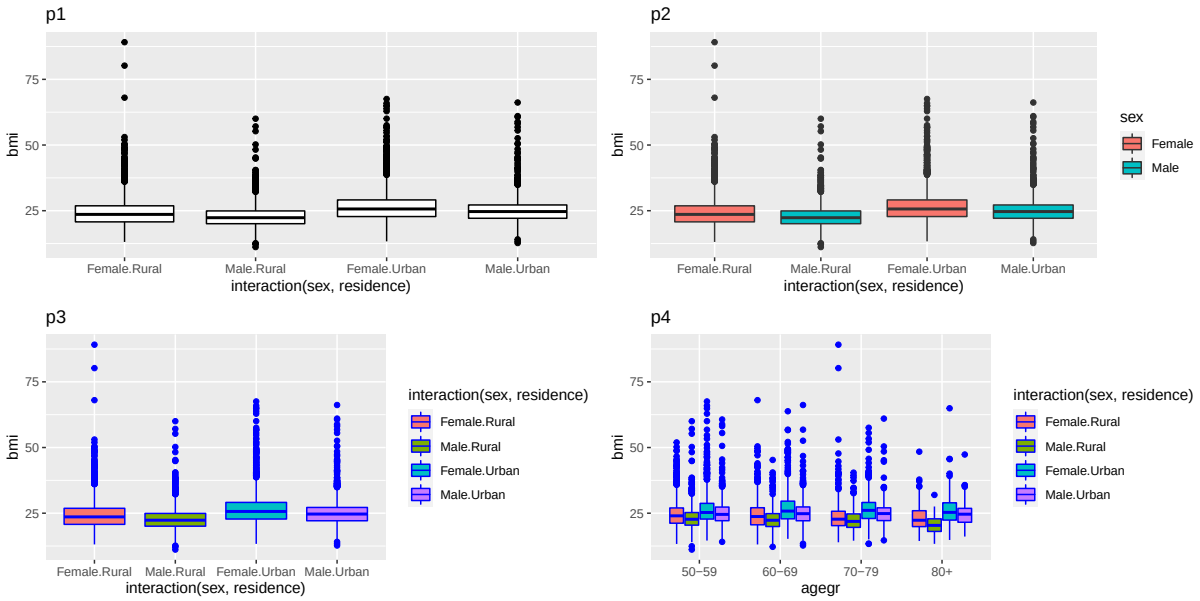

Figure 1.7: Boxplot; examples of using interaction and aes

1.3 Using facets

This facility makes it easy to split the output into different plots by categorical data using the formula object.

```
p<-ggplot(norsjo86,aes(y=sbp,x=bmi))+
geom_point(size=1.5)
p1<-p+facet_grid(.~as_factor(sex))
p2<-p+facet_grid(as_factor(sex)~agegrp)
```

```
p1
```

```
## Warning:  Removed 5 rows containing missing values (geom_point).
```

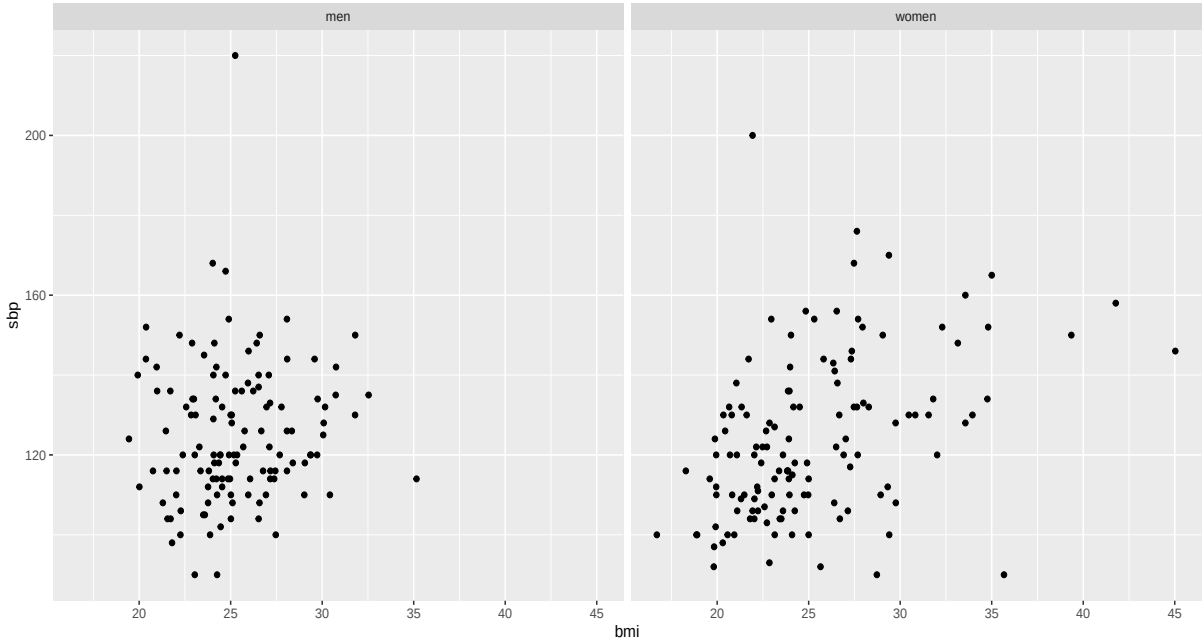

Figure 1.8: Scatterplots of sbp vs bmi split by sex

```
p2
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

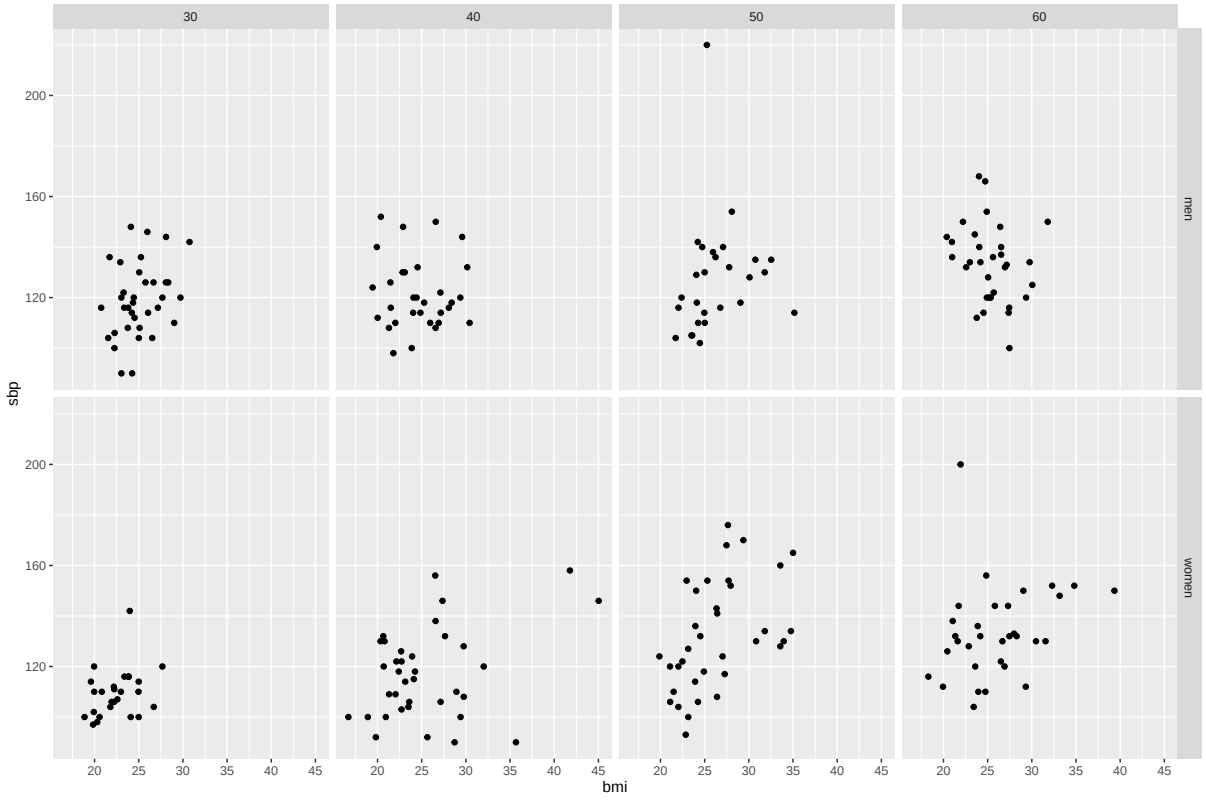

Figure 1.9: Scatterplots of sbp vs bmi split by sex and age

```
p<-ggplot(norsjo86,aes(y=sbp,x=bmi))+
geom_point(size=1.5)+
facet_grid(as_factor(sex)~agegrp)

ps1<-p+geom_smooth(method=lm,se=T)+ # Add smoothing with linear regression
labs(title="ps1")
ps2<-p+geom_smooth(se=F)+ # by default it is smoothed using loess
labs(title="ps2")
ps3<-p+geom_smooth(se=F,span=0.4)+  # less stiff
labs(title="ps3")
ps4<-p+geom_smooth(se=F,span=10,color="black")+ # more stiff
labs(title="ps4")
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
## Warning: Removed 5 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
## Warning: Removed 5 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
## Warning: Removed 5 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
## Warning: Removed 5 rows containing missing values (geom_point).
```

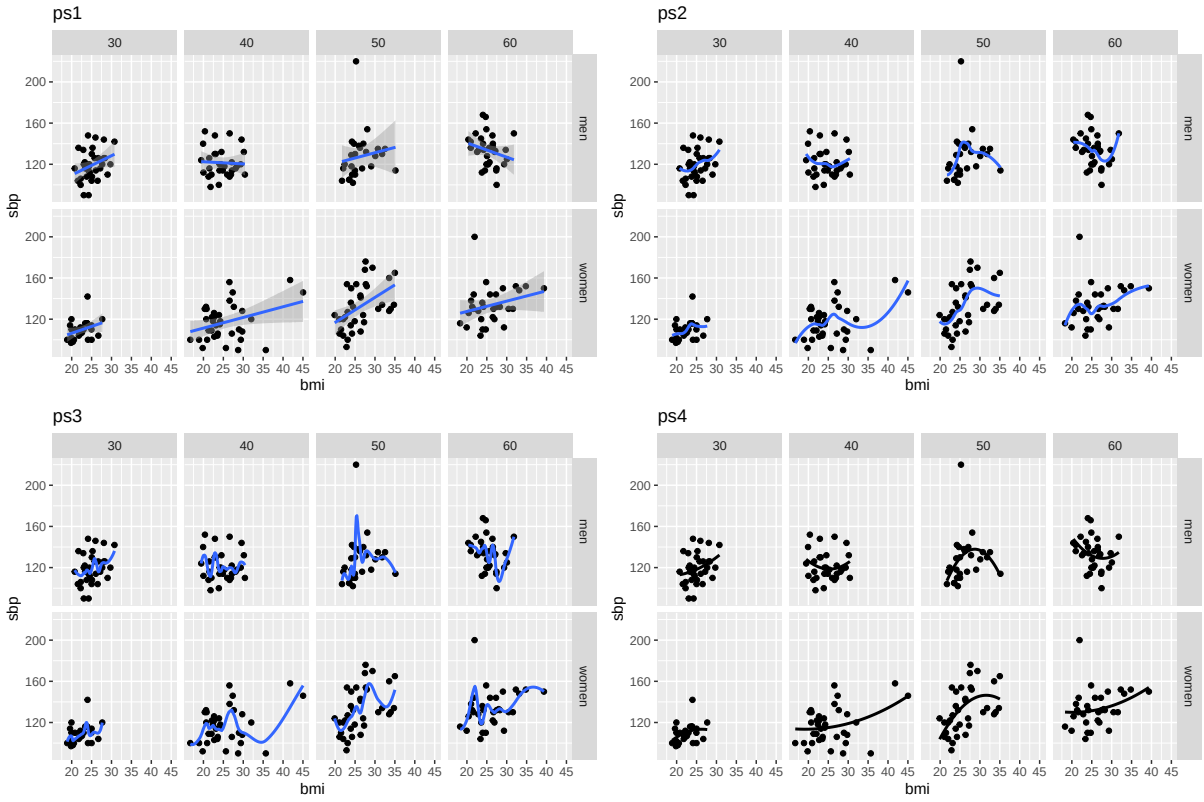

Figure 1.10: Scatterplots with facets and smoothing

```
p<-ggplot(norsjo86,aes(y=sbp,x=bmi))+
  facet_grid(as_factor(sex)~agegrp)
pf1<-p+geom_point(size=2,shape=21,aes(fill=as_factor(smoker)))+
   labs(title="pf1")
pf2<-p+geom_point(size=2,aes(color=cut(cholesterol,c(0,6,100)),shape=as_factor(smoker)))+
   labs(title="pf2")
```

We could have chosen to first add the factors of cholesterol and smoker to the data frame to get a more nice legend.

```
## Warning:  Removed 5 rows containing missing values (geom_point).
## Error in cut(cholesterol, c(0, 6, 100)):  object 'cholesterol' not found
```

Figure 1.11: Scatterplots with facets and aes

1.3.1 Using facets and other grouping together with converting data into long format

We first use the tidyr::gather function to store the data into long format. The data is duplicated for sex and agegr and the variable "var" (key) keeps track on which variable represents the the "value". The key and value variables are used in aes.

```
bp.lf<- bp %>% gather(key=var,value=value,sex,agegr)

bp.lf %>% select(id,disability, residence,country,var,value) %>% head(5)

Error in select(., id, disability, residence, country, var, value):  unused arguments (id,
disability, residence, country, var, value)

bp.lf %>% select(id,disability, residence,country,var,value) %>% tail(5)

Error in select(., id, disability, residence, country, var, value):  unused arguments (id,
disability, residence, country, var, value)

table(bp.lf$var)

agegr   sex
10000 10000

pb1<-ggplot(bp.lf,aes(y=disability,x=value,fill=var))+
geom_boxplot(colour="blue",size=0.5)+ # disability is plotted vs both sex and agegr
labs(title="pb1")

pb2<-pb1+facet_grid(residence~.)+       # split by residence
labs(title="pb2")
pb3<-pb1+facet_grid(residence~country)+ # also split by country
labs(title="pb3")
```

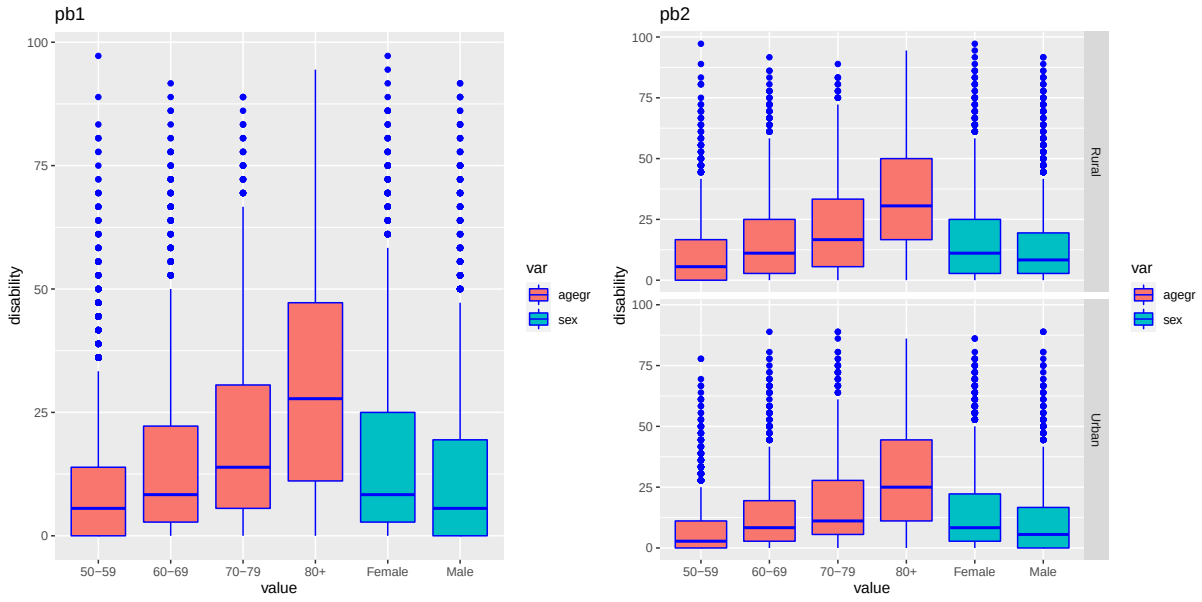

Figure 1.12: Boxplots using long format (two variables in one)

So far we have used **gather** on factors to use these in aes options. However, we can also gather on continuous variables. This gives a possibility to plot different variables in the same plot.

```
pb3
```

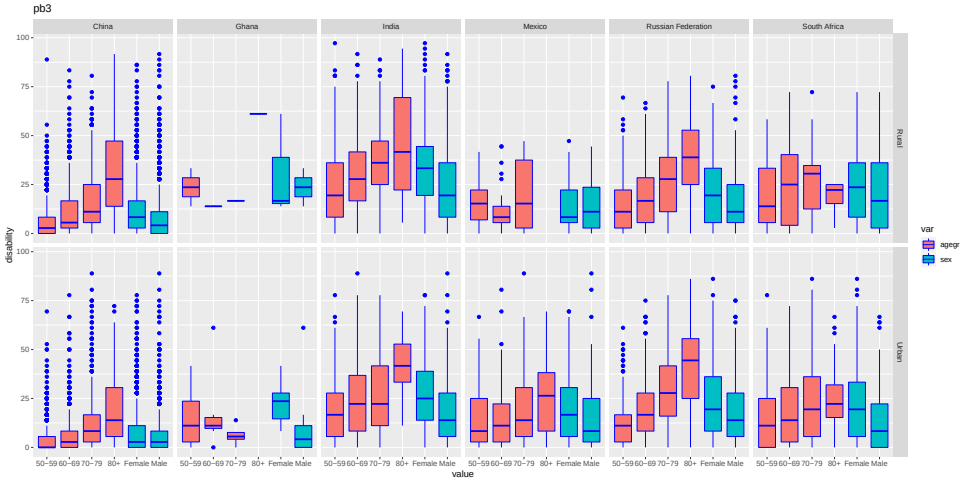

Figure 1.13: Boxplots using long format and facets

```
# this time we gather continuous variables
bp.lf<- bp %>% gather(key=var,value=value,bmi,height)

pc1<-ggplot(bp.lf,aes(x=waistc,y=value,fill=var))+
geom_point(size=1.3,shape=21)

pc2<-pc1+facet_grid(residence~physical)+
   geom_smooth(method=lm,se=F,color="black")
```

```
pc1
```

Figure 1.14: Plot of two continuous variables in one using gather

```
pc2
```

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

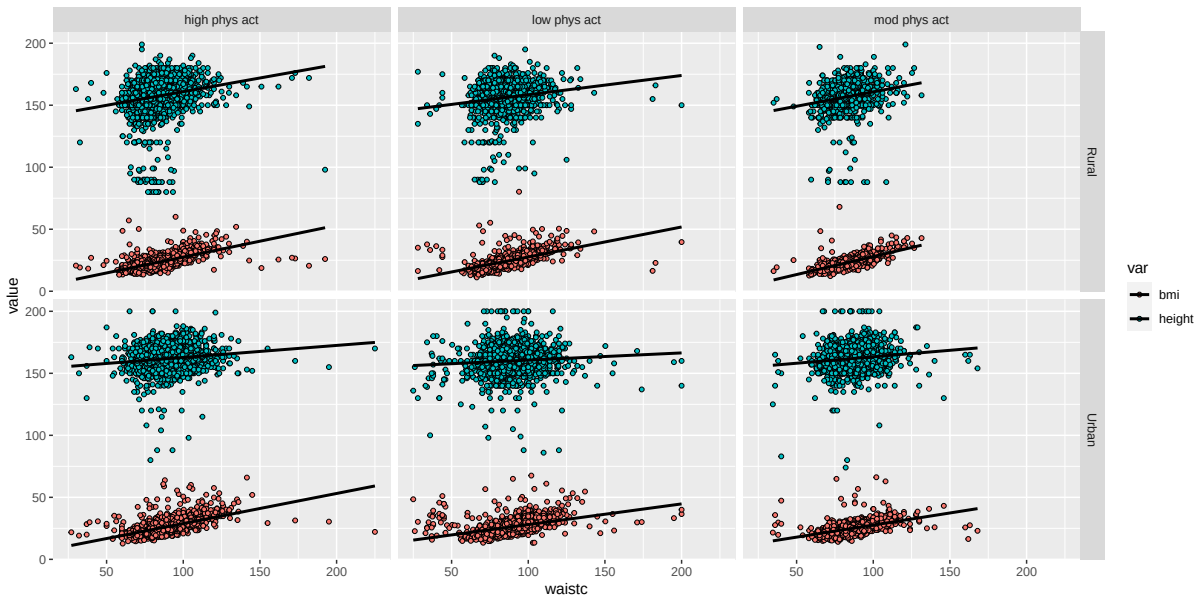

Figure 1.15: Plot of two continuous variables in one using gather and facets