STAT 224 Lecture 1
R and Intro to ggplot2

Yibi Huang

Outline

- The Lung Capacity Data
- Review of Introductory Statistics
- Summarizing data *by group* using `aggregate()` in R
- The `ggplot` library in R

The FEV Lung Capacity Data

Data: Smoking & FEV (Lung Capacity)

Sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970's. The variables are

- **age**: Subject's age in years
- **fev**: Lung capacity of subject, measured by **forced expiratory volume** (abbreviated as **FEV**), the amount of air an individual can exhale in the first second of forceful breath in liters
- **ht**: Subject's height in inches
- **sex**: Gender of the subject coded as: 0 = Female, 1 = Male
- **smoke**:Smoking status coded as: 0 = Nonsmoker, 1 = Smoker

Loading Data

You can download the data at

`http://www.stat.uchicago.edu/~yibi/s224/data/fevdata.txt`

save it on your computer, and then **change the working directory** to the folder the file `fevdata.txt` is saved. (See Slides L00).

You can load the data to R by the commands below.

```
fevdata = read.table("fevdata.txt", header = TRUE)
```

The first few rows of the data:

```
head(fevdata)
  age   fev   ht sex smoke
1   9 1.708 57.0   0     0
2   8 1.724 67.5   0     0
3   7 1.720 54.5   0     0
4   9 1.558 53.0   1     0
5   9 1.895 57.0   1     0
```

Finding Data Summaries in R

Finding the Sample Mean in R

```
mean(fev)
Error in mean(fev): object 'fev' not found
```

Oops! Error message.

R cannot find the variable `fev` unless you indicate it's inside the data frame `fevdata` as follows.

```
mean(fevdata$fev)
[1] 2.637
```

Other Summary Statistics

```
median(fevdata$fev)
[1] 2.547
sd(fevdata$fev)
[1] 0.8671
var(fevdata$fev)
[1] 0.7518
```

Also works for `max()`, `min()`, `sum()`, `IQR()`.

Data Summaries "By Group"

Other than summarizing data over the **entire** data set, one might want to summarize data (mean, median, SD, IQR, etc) **by group**, e.g., by gender, by smoking status, or by some other grouping variables.

Sex/Smoke	Count	Mean	SD
Female Nonsmoker	279	2.379	0.639
Female Smoker	39	2.966	0.423
Male Nonsmoker	310	2.734	0.974
Male Smoker	26	3.743	0.889

Summarizing Data "By Group" Using `aggregate()`

The `aggregate()` function in R allows users to summarize data **by group** with simple syntax.

Mean FEV by Gender: (0 for females, 1 for males)

```
aggregate(fev ~ sex, data = fevdata, mean)
  sex   fev
1   0 2.451
2   1 2.812
```

Mean FEV by Smoking Status: (0 for nonsmokers, 1 for smokers)

```
aggregate(fev ~ sex, data = fevdata, mean)
  sex   fev
1   0 2.451
2   1 2.812
```

Changing the Labels of Categorical Variables

Better giving **meaningful labels** for the categories of `sex` and `smoke` rather than 0 and 1.

```
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

```
aggregate(fev ~ sex, data = fevdata, mean)
```

```
     sex   fev
1 Female 2.451
2   Male 2.812
```

Do males or females have a greater mean FEV (lung capacity)?

```
aggregate(fev ~ smoke, data = fevdata, mean)
```

```
      smoke   fev
1 Nonsmoker 2.566
2    Smoker 3.277
```

Do smokers or nonsmokers have a greater mean FEV?

Changing the Labels of Categorical Variables

Better giving **meaningful labels** for the categories of `sex` and `smoke` rather than 0 and 1.

```
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

```
aggregate(fev ~ sex, data = fevdata, mean)
```

```
     sex   fev
1 Female 2.451
2   Male 2.812
```

Do males or females have a greater mean FEV (lung capacity)?

```
aggregate(fev ~ smoke, data = fevdata, mean)
```

```
      smoke   fev
1 Nonsmoker 2.566
2    Smoker 3.277
```

Do smokers or nonsmokers have a greater mean FEV?

Surprising! How could smokers have a greater mean lung capacities than nonsmokers?

Changing the Labels of Categorical Variables

Better giving **meaningful labels** for the categories of `sex` and `smoke` rather than 0 and 1.

```
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

```
aggregate(fev ~ sex, data = fevdata, mean)
```

```
     sex   fev
1 Female 2.451
2   Male 2.812
```

Do males or females have a greater mean FEV (lung capacity)?

```
aggregate(fev ~ smoke, data = fevdata, mean)
```

```
      smoke   fev
1 Nonsmoker 2.566
2    Smoker 3.277
```

Do smokers or nonsmokers have a greater mean FEV?

Surprising! How could smokers have a greater mean lung capacities than nonsmokers?

Summarizing Data By Two Grouping Variables

Mean FEV by both gender & smoking status:

```
aggregate(fev ~ smoke + sex, data = fevdata, mean)
      smoke    sex   fev
1 Nonsmoker Female 2.379
2    Smoker Female 2.966
3 Nonsmoker   Male 2.734
4    Smoker   Male 3.743
```

Summarizing Data By Two Grouping Variables

Mean FEV by both gender & smoking status:

```
aggregate(fev ~ smoke + sex, data = fevdata, mean)
      smoke    sex   fev
1 Nonsmoker Female 2.379
2    Smoker Female 2.966
3 Nonsmoker   Male 2.734
4    Smoker   Male 3.743
```

- For females, did smokers or non-smokers have greater mean lung capacities?
- For males, did smokers or non-smokers have greater mean lung capacities?

Other Summary Statistics "by Group"

The `aggregate()` function can find other summary statistics: `median()`, `sd()`, `var()`, `max()`, `min()`, `sum()`, etc, **by group** following the syntax:

```
aggregate(variable ~ grouping_var, data=dataframename, stat_fun)
```

```
aggregate(fev ~ smoke, data = fevdata, sd)
      smoke    fev
1 Nonsmoker 0.8505
2    Smoker 0.7500
aggregate(fev ~ smoke + sex, data = fevdata, median)
      smoke    sex   fev
1 Nonsmoker Female 2.417
2    Smoker Female 3.074
3 Nonsmoker   Male 2.547
4    Smoker   Male 3.878
```

aggregate() + summary()

The summary() function can report the five-number summary + mean.

```
aggregate(fev ~ smoke + sex, data=fevdata, summary)
      smoke    sex fev.Min. fev.1st Qu. fev.Median fev.Mean fev.3rd Qu.
1 Nonsmoker Female    0.791       1.877      2.417    2.379       2.865
2    Smoker Female    2.198       2.678      3.074    2.966       3.197
3 Nonsmoker   Male    0.796       1.964      2.547    2.734       3.353
4    Smoker   Male    1.694       3.359      3.878    3.743       4.380
  fev.Max.
1    3.816
2    3.835
3    5.793
4    4.872
```

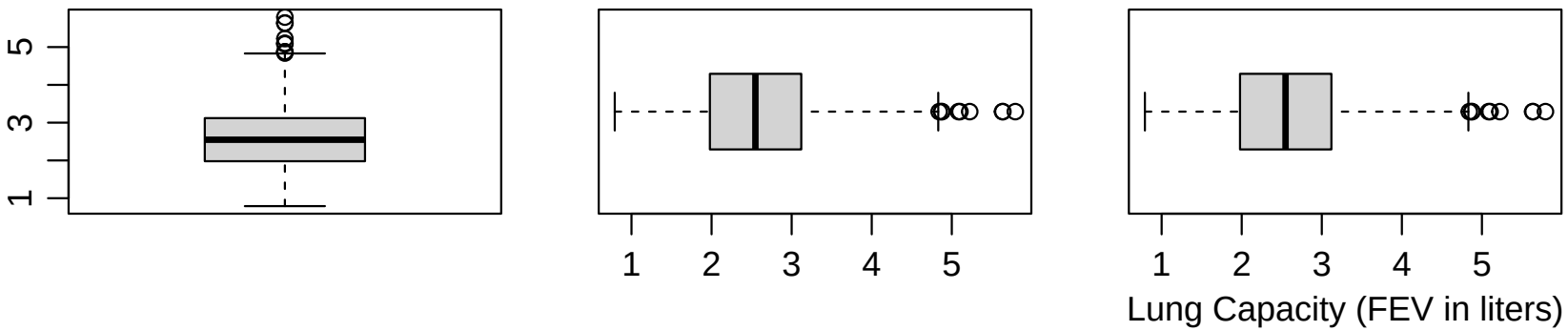
Do smokers still appear to have greater lung capacities than nonsmokers?

Graphical Summary and the `ggplot2` Library

Boxplots

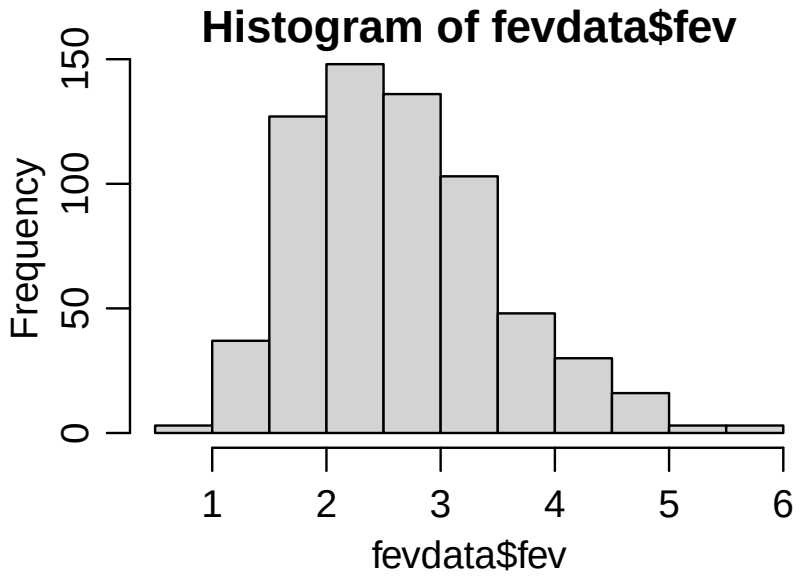
```
summary(fevdata$fev)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.791   1.981   2.547   2.637   3.118   5.793
```

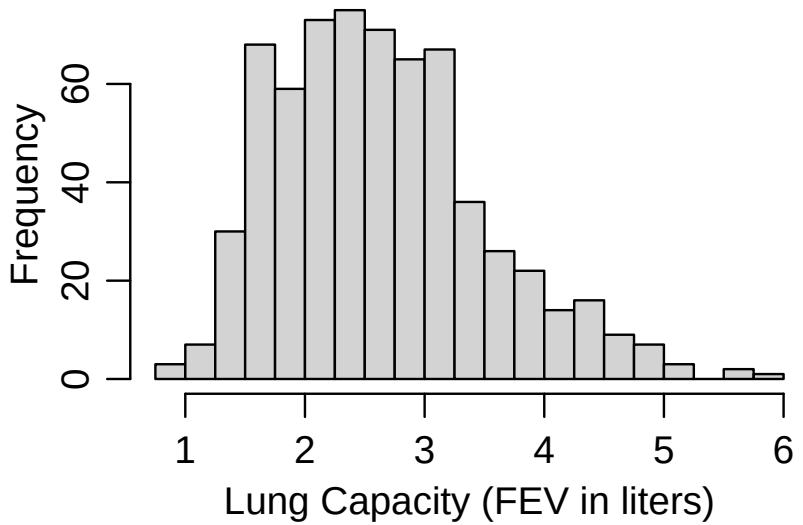
```
boxplot(fevdata$fev)
boxplot(fevdata$fev, horizontal = T)
boxplot(fevdata$fev, horizontal = T,
        xlab="Lung Capacity (FEV in liters)")
```



Histograms

```
hist(fevdata$fev)
hist(fevdata$fev, breaks=seq(0.75,6, 0.25),
          xlab="Lung Capacity (FEV in liters)", main="")
```





The `ggplot2` Library

- “gg” in `ggplot` means = **G**rammer of **G**raphics

The `ggplot2` Library

- "gg" in `ggplot` means = **G**rammer of **G**raphics
- a powerful and versatile toolkit for data visualization

The ggplot2 Library

- "gg" in `ggplot` means = **G**rammer of **G**raphics
- a powerful and versatile toolkit for data visualization

Before installing the `ggplot2` library, make sure you haven't installed it by loading the library.

```
library(ggplot2)
```

If you get no message, that means the `ggplot2` has been installed on your machine. You can move on and skip installation .

If you get the error message below

```
# Error in library(ggplot2) : there is no package called 'ggplot2'
```

that means the `ggplot2` library is not installed. You can install it using the command below.

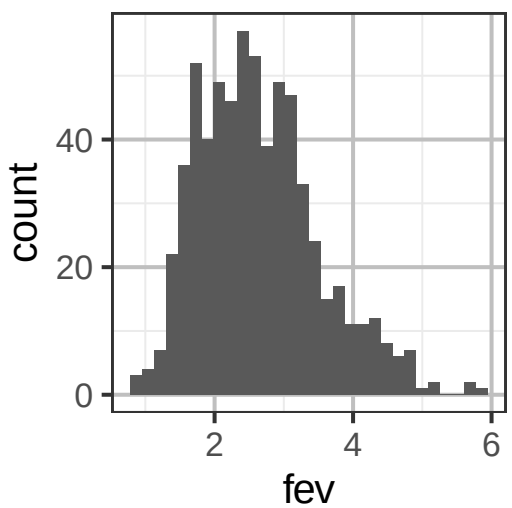
```
install.packages("ggplot2")
```

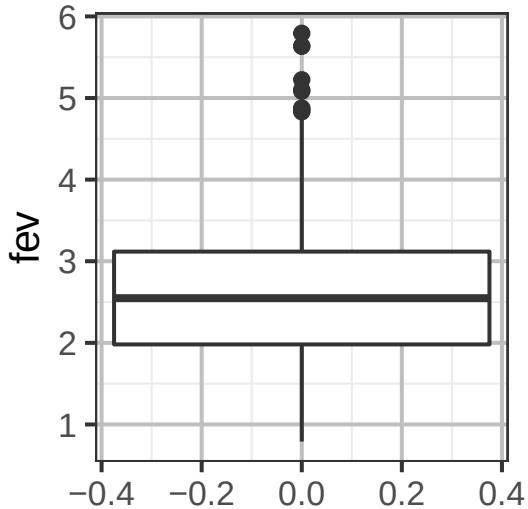
Histograms, Boxplots, Scatterplots in ggplot2

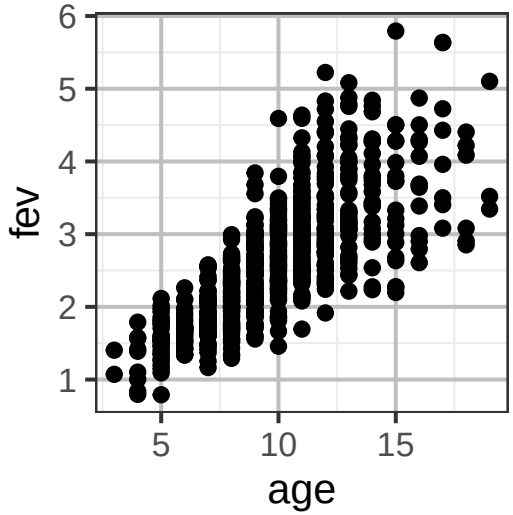
`aes()` is the short hand for “aesthetic”.

Variables involved in a ggplot must be enclosed within `aes()`.

```
ggplot(fevdata, aes(x = fev)) + geom_histogram()
ggplot(fevdata, aes(y = fev)) + geom_boxplot()
ggplot(fevdata, aes(x = age, y = fev)) + geom_point()
```

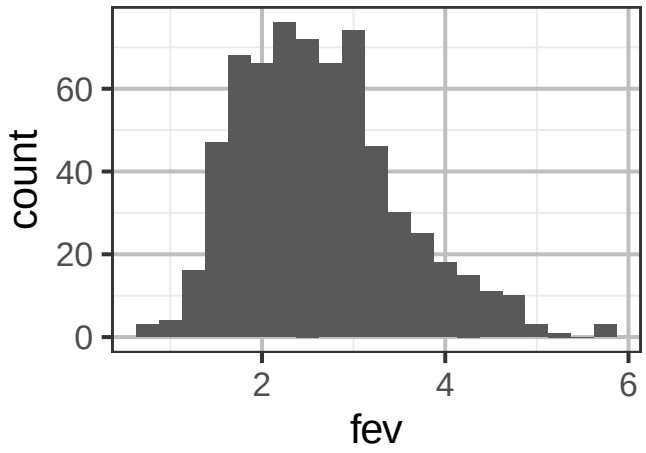




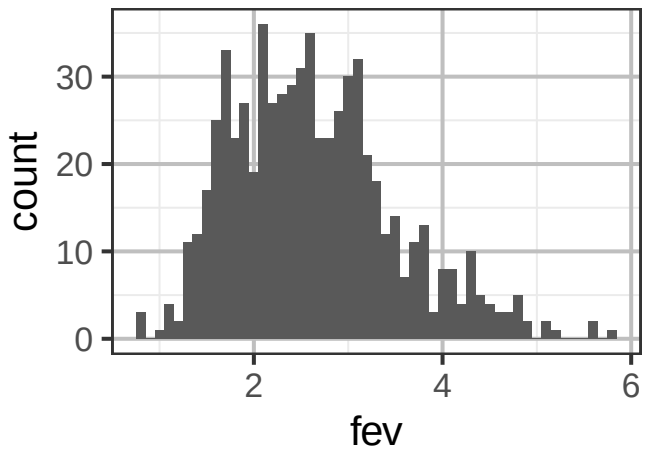


Please ALWAYS Adjust the Binwidth of Histograms

```
ggplot(fevdata, aes(x = fev)) + geom_histogram(binwidth = 0.25)
```



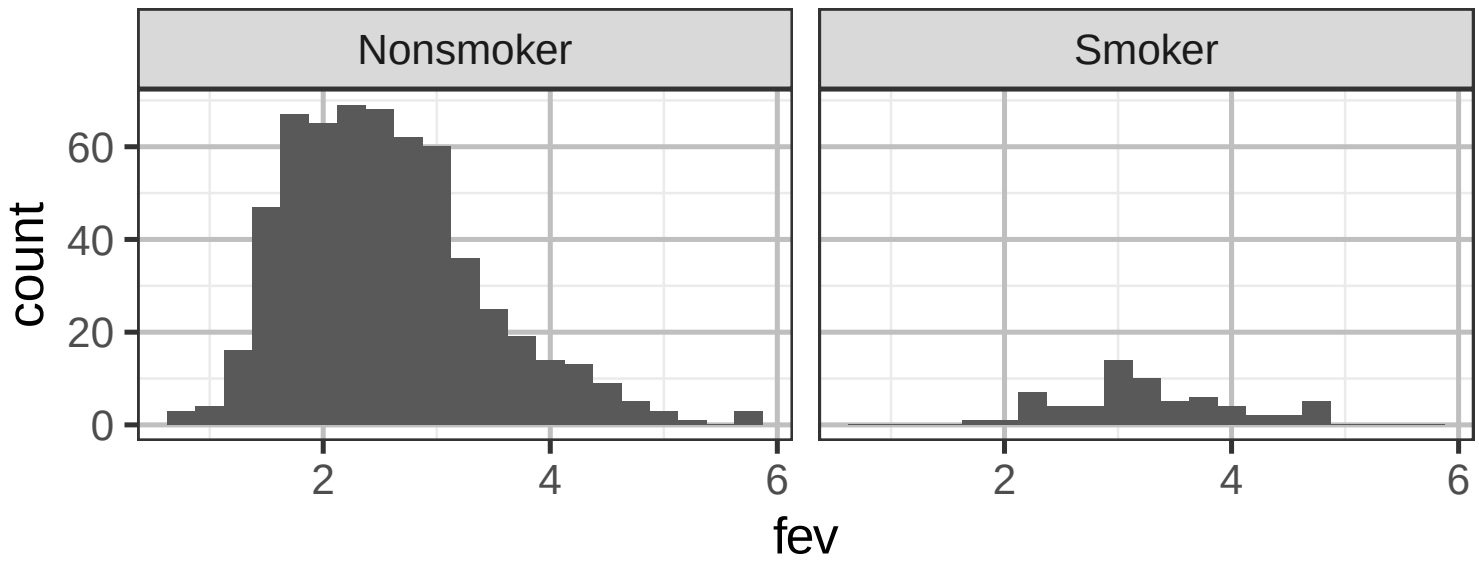

```
ggplot(fevdata, aes(x = fev)) + geom_histogram(binwidth = 0.1)
```



Facet — Inspect Data by Group

`ggplot` allows you to inspect data by group using `facet_wrap`.

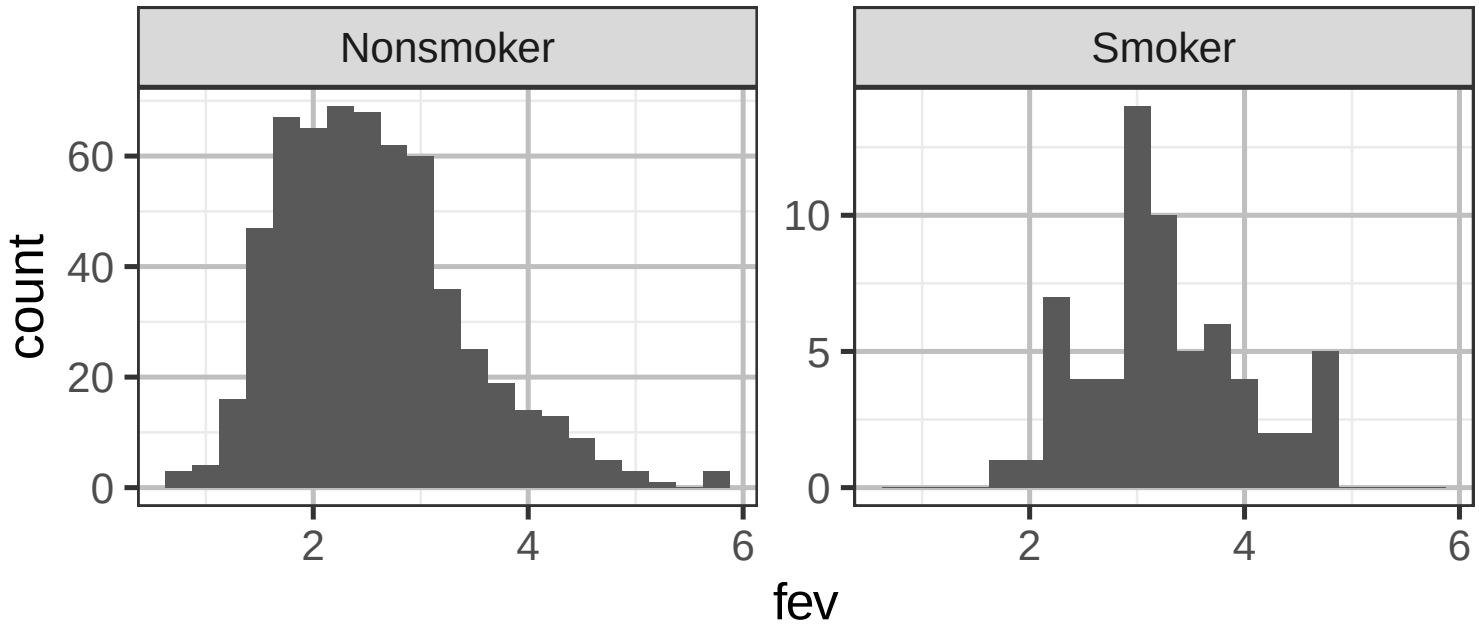
```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(~smoke)
```



The histogram for nonsmokers is taller than the one for smokers since there are far more nonsmokers than smokers in the data and the two histograms **share the same y-axis**.

You can free up the y axis so they can be different between the graphs by adding `scale="free_y"` within `facet_wrap()`.

```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(~smoke, scale="free_y")
```

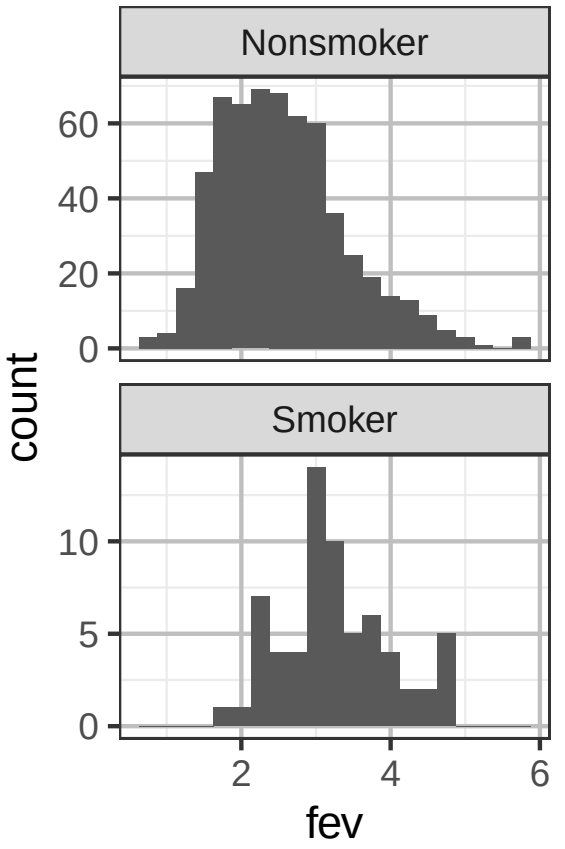


Stacking Histograms Vertically

Stacking histograms vertically makes it easier to **compare the horizontal scale**.

Adding `nrow=2` in `facet_wrap` arranges the histograms in 2 rows.

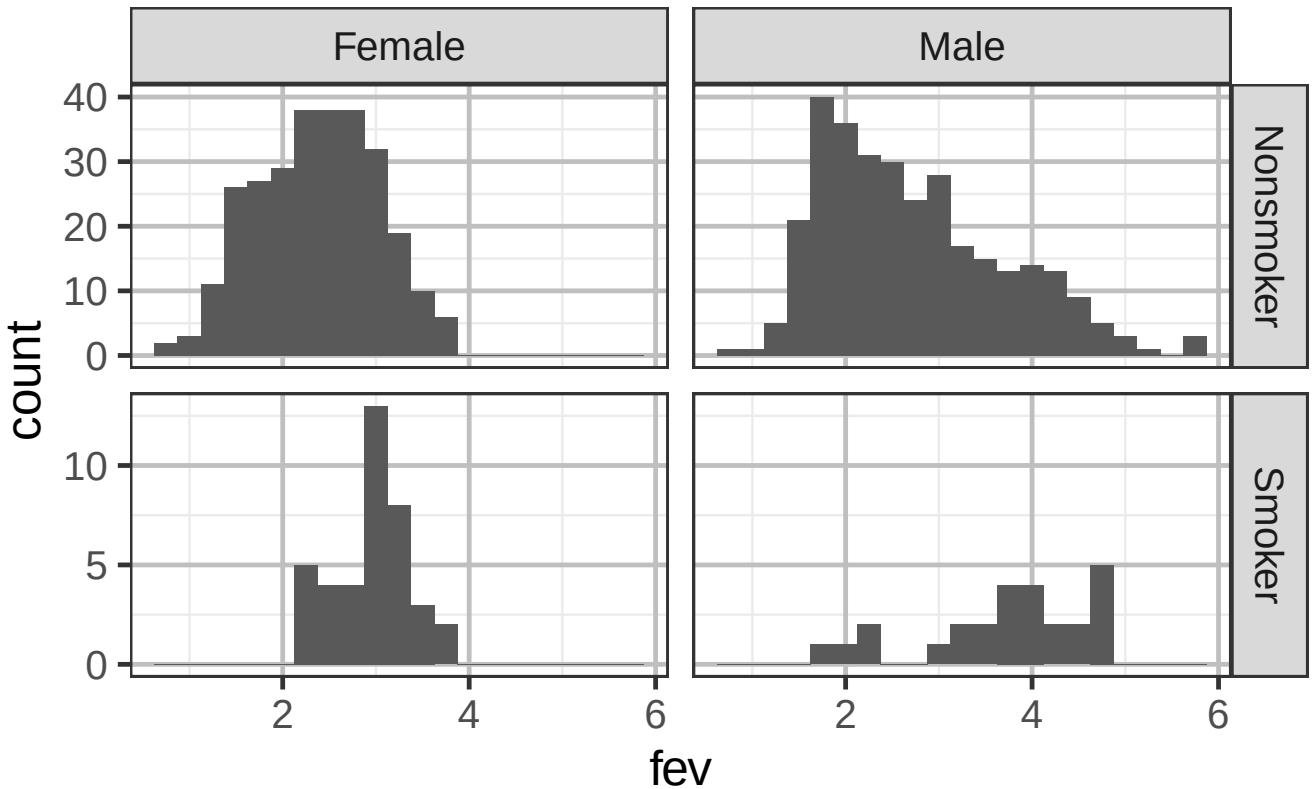
```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(~smoke,scale="free_y",nrow=2)
```



Do nonsmokers have greater lung capacity than smokers?

Facet Over 2 Grouping Variables — facet_grid

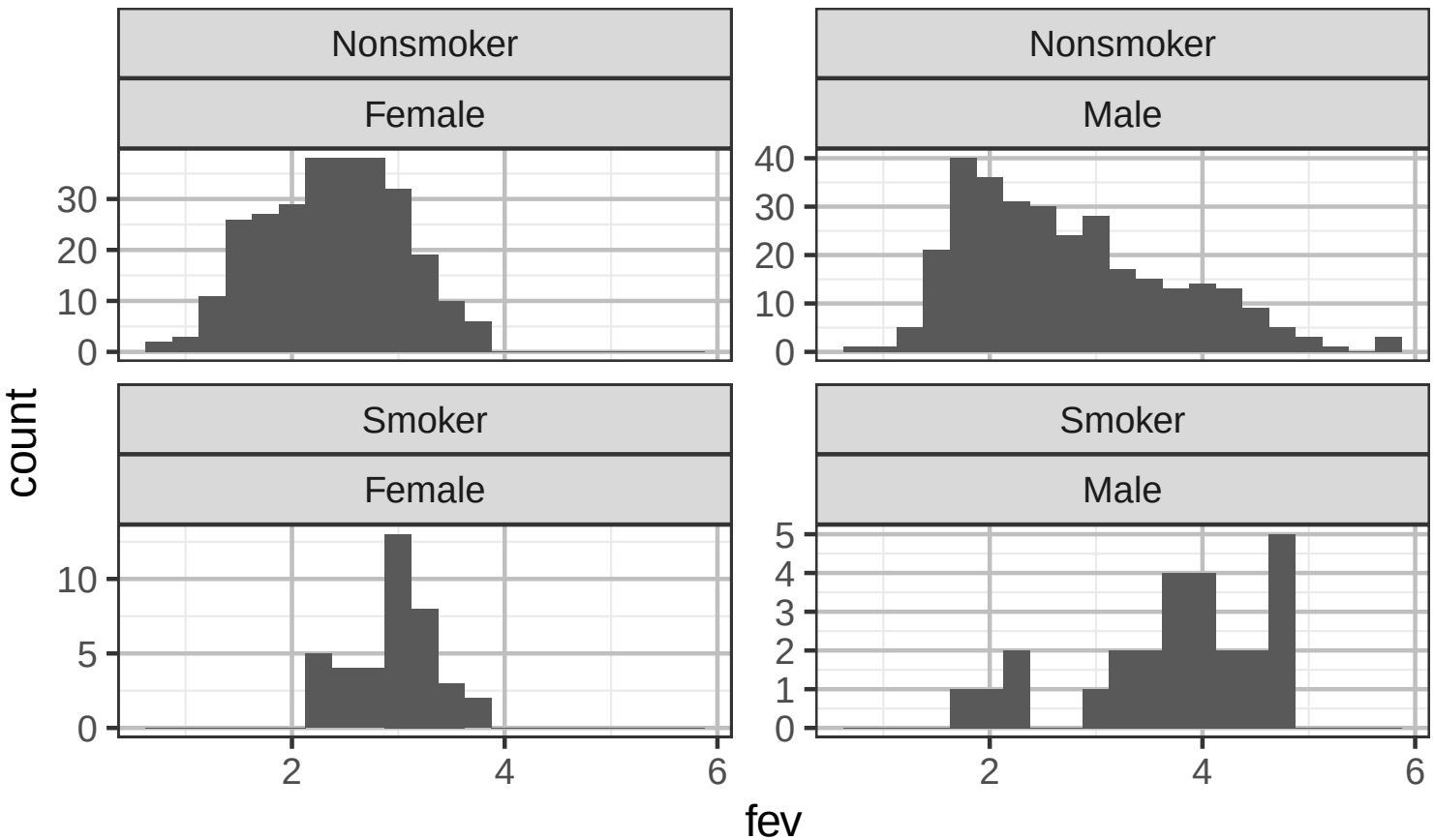
```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_grid(smoke ~ sex, scale="free_y")
```



For females, did smokers or nonsmokers have greater lung capacity? How about males?

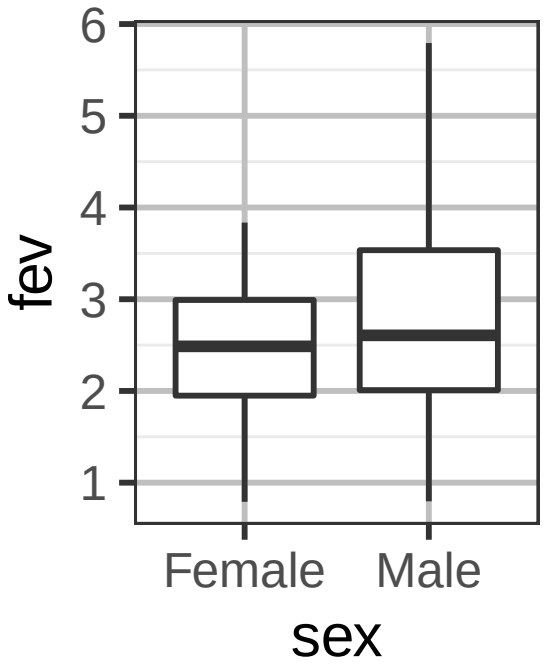
One can also facet over 2 grouping variables using `facet_wrap()` but the 2 layers of grouping labels take too much space.

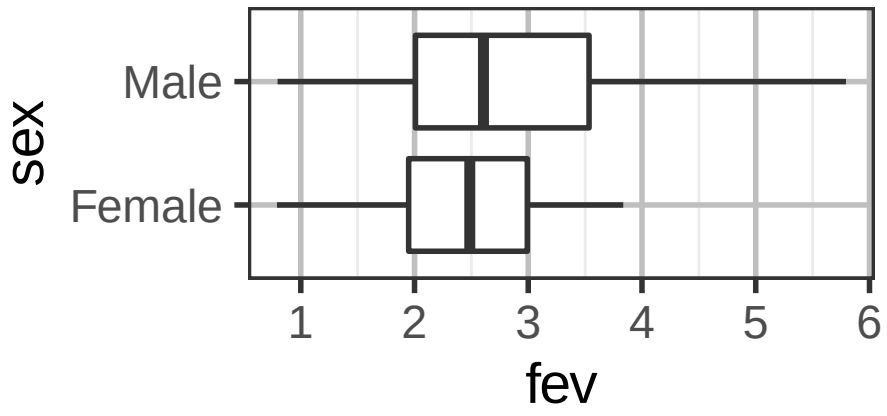
```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(smoke ~ sex, scale="free_y", nrow=2)
```



Boxplots by Group (or Side-by-Side Boxplots)

```
ggplot(fevdata, aes(x = sex, y = fev)) + geom_boxplot()
ggplot(fevdata, aes(x= fev, y = sex)) + geom_boxplot()
```

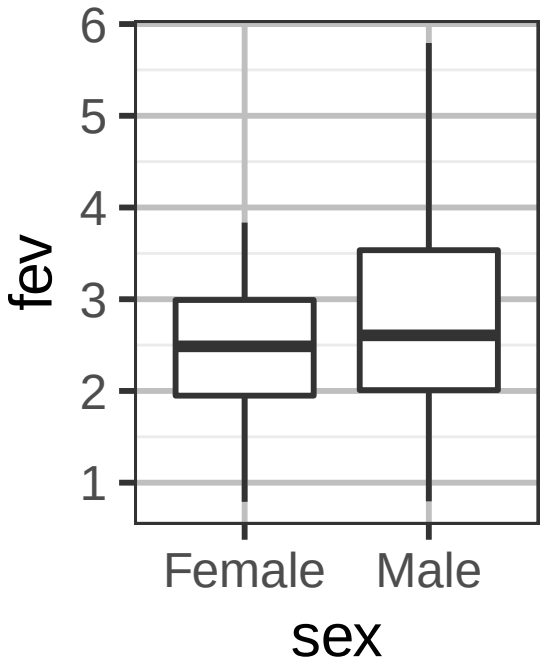


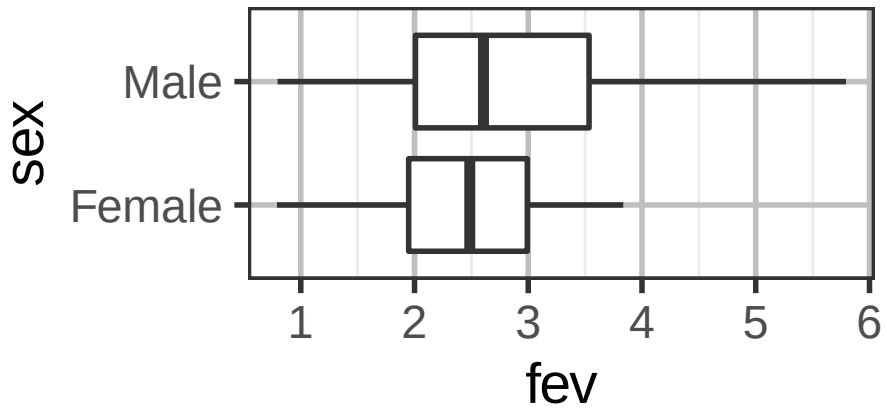


- Boxplots can be made *horizontal* by flipping x & y

Boxplots by Group (or Side-by-Side Boxplots)

```
ggplot(fevdata, aes(x = sex, y = fev)) + geom_boxplot()
ggplot(fevdata, aes(x= fev, y = sex)) + geom_boxplot()
```

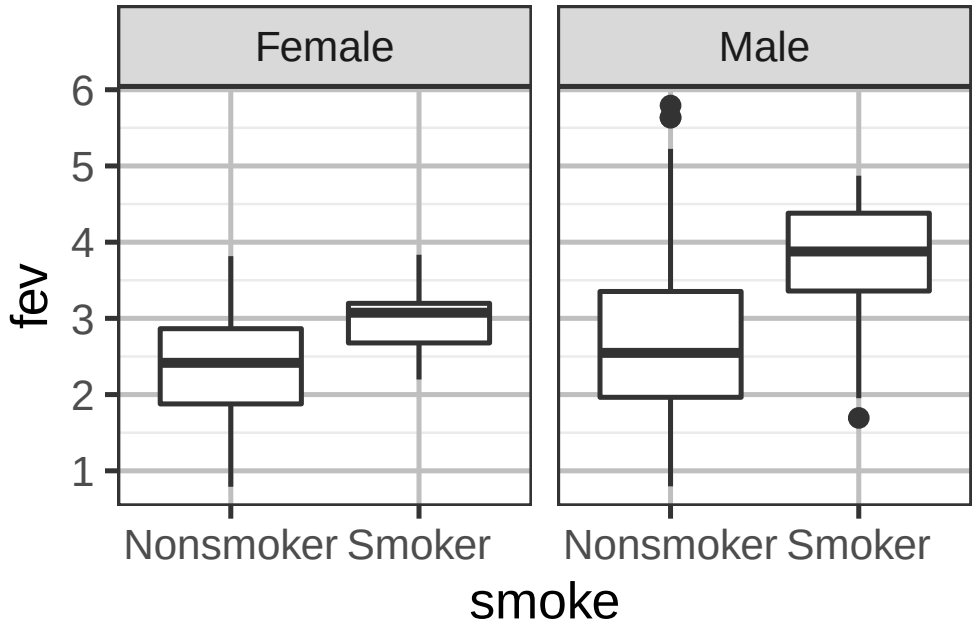




- Boxplots can be made *horizontal* by flipping x & y
- Did males or females have greater lung capacity?

Faceting Boxplots

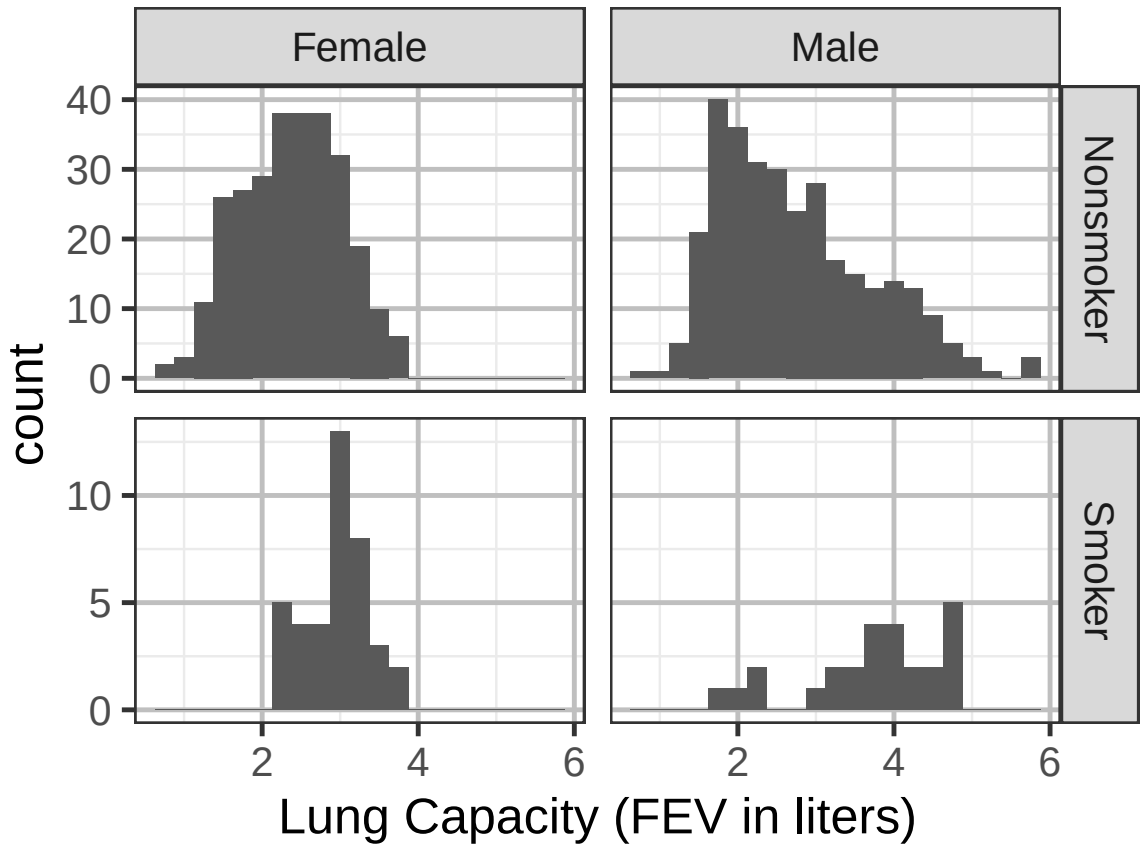
```
ggplot(fevdata, aes(x = smoke, y = fev)) +
  geom_boxplot() +
  facet_wrap(~sex)
```



For females, do smokers or nonsmokers have greater lung capacities? How about males?

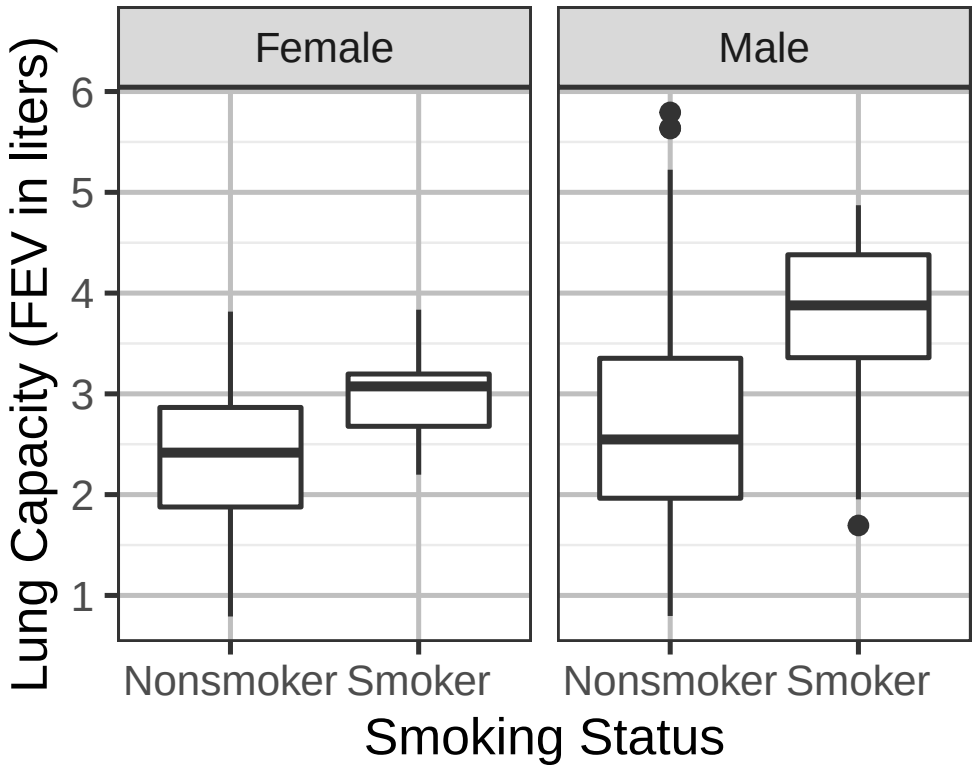
Please Label Your Plots Properly

```
ggplot(fevdata, aes(x = fev)) +
  geom_histogram(binwidth = 0.25) +
  facet_grid(smoke~sex, scale="free_y") +
  xlab("Lung Capacity (FEV in liters)")
```



Please Label Your Plots Properly

```
ggplot(fevdata, aes(x=smoke, y=fev)) +
  geom_boxplot() +
  facet_wrap(~sex) +
  xlab("Smoking Status") +
  ylab("Lung Capacity (FEV in liters)")
```



T-test – FEV of Smokers & Nonsmokers

Do smokers have a significantly higher mean FEV than nonsmokers?

```
t.test(fev ~ smoke, data=fevdata)

    Welch Two Sample t-test

data:  fev by smoke
t = -7.1, df = 83, p-value = 0.0000000003
alternative hypothesis: true difference in means between group Nonsmoke
95 percent confidence interval:
 -0.9084 -0.5130
sample estimates:
mean in group Nonsmoker    mean in group Smoker
                  2.566                   3.277
```

The broom Library Can Tidy Up R Output

The `tidy()` function in the broom library can give a tidier output than the verbose R output for `t.test()`.

```
library(broom)
tidy(t.test(fev ~ smoke, data=fevdata))
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic  p.value parameter conf.low co
     <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1   -0.711      2.57      3.28     -7.15 3.07e-10      83.3   -0.908
# ... with 2 more variables: method <chr>, alternative <chr>
```

Need to install the *broom* library if you haven't just like the ggplot2 library.

```
install.packages("broom")
```

Sample v.s. Population

- Which of the following is the correct null hypothesis for the t-test on the previous page, assuming that the data were randomly sampled from some population?

1. Smokers in the sample had an identical mean FEV as nonsmokers in the sample.
2. Smokers in the population had an identical mean FEV as nonsmokers in the population.

Sample v.s. Population

- Which of the following is the correct null hypothesis for the t-test on the previous page, assuming that the data were randomly sampled from some population?

1. Smokers in the sample had an identical mean FEV as nonsmokers in the sample.
2. Smokers in the population had an identical mean FEV as nonsmokers in the population.

- Keep in mind that we are interested in the population, not the sample.

Sample v.s. Population

- Which of the following is the correct null hypothesis for the t-test on the previous page, assuming that the data were randomly sampled from some population?

1. Smokers in the sample had an identical mean FEV as nonsmokers in the sample.
2. Smokers in the population had an identical mean FEV as nonsmokers in the population.

- Keep in mind that we are interested in the population, not the sample.

- Interpretation of the 95% confidence interval $(-0.908, -0.513)$: With 95% confidnece, the mean FEV of nonsmokers is 0.908 to 0.513 liters lower than that of smokers.

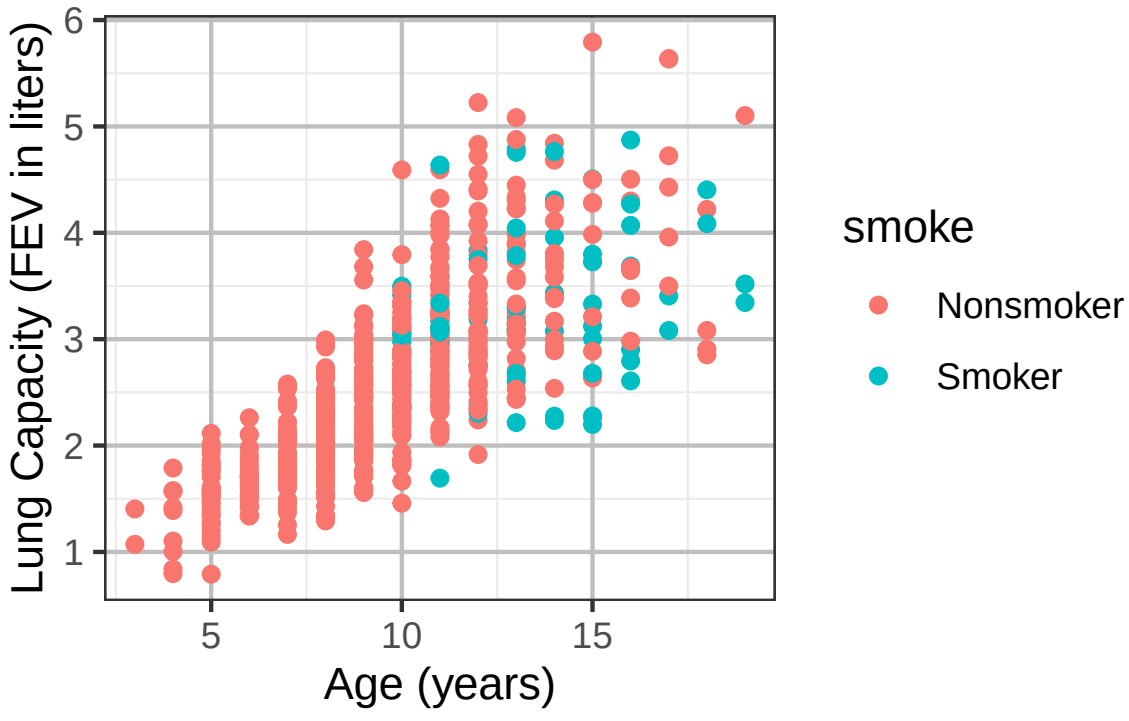
T-test — FEV of Smokers & Nonsmokers by Sex

Do smokers have a significantly higher mean FEV than nonsmokers **of the same gender**?

```
tidy(t.test(fev ~ smoke, data=subset(fevdata, sex=="Male")))
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic    p.value parameter conf.low
     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>     <dbl>    <dbl>
1    -1.01      2.73      3.74     -5.51 0.00000533      30.3    -1.38
# ... with 2 more variables: method <chr>, alternative <chr>
tidy(t.test(fev ~ smoke, data=subset(fevdata, sex=="Female")))
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic  p.value parameter conf.low co
     <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1   -0.587      2.38      2.97     -7.54 1.85e-10      65.2   -0.742
# ... with 2 more variables: method <chr>, alternative <chr>
```

Coded Scatter Plots

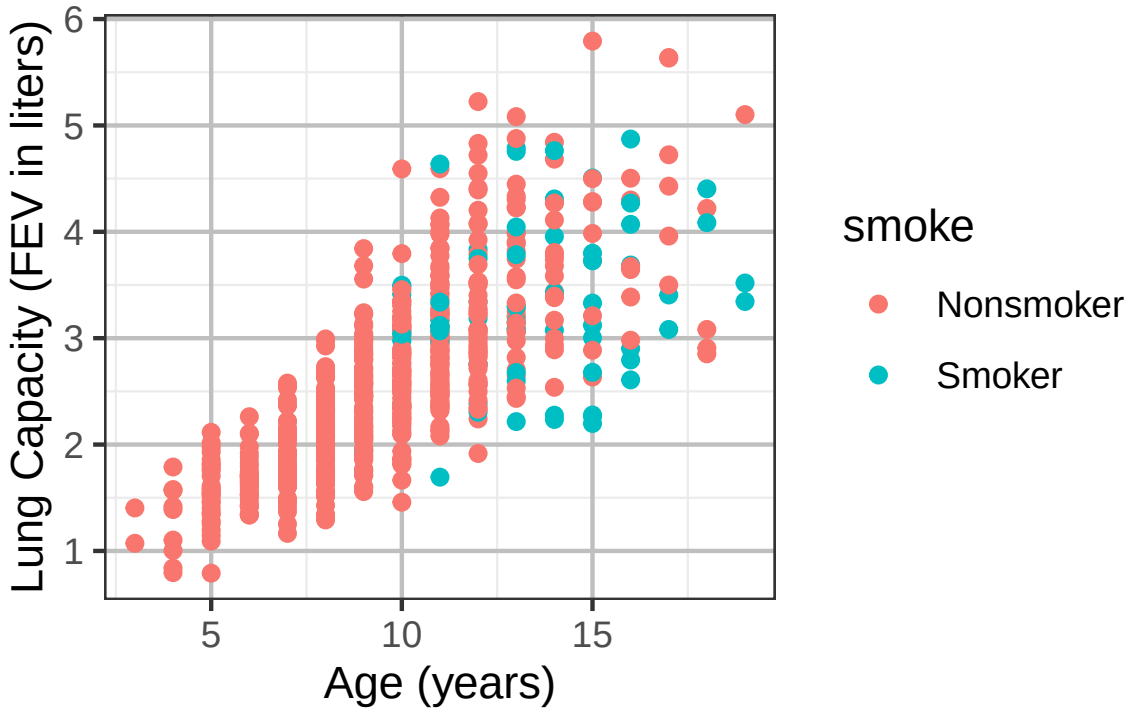
```
ggplot(fevdata, aes(x = age, y = fev, col=smoke)) + geom_point() +
  xlab("Age (years)") + ylab("Lung Capacity (FEV in liters)")
```



- In `ggplot`, it's easy to make coded scatter plots that the **color** or **shape** of points represent a third variable, like `smoke`

Coded Scatter Plots

```
ggplot(fevdata, aes(x = age, y = fev, col=smoke)) + geom_point() +
  xlab("Age (years)") + ylab("Lung Capacity (FEV in liters)")
```

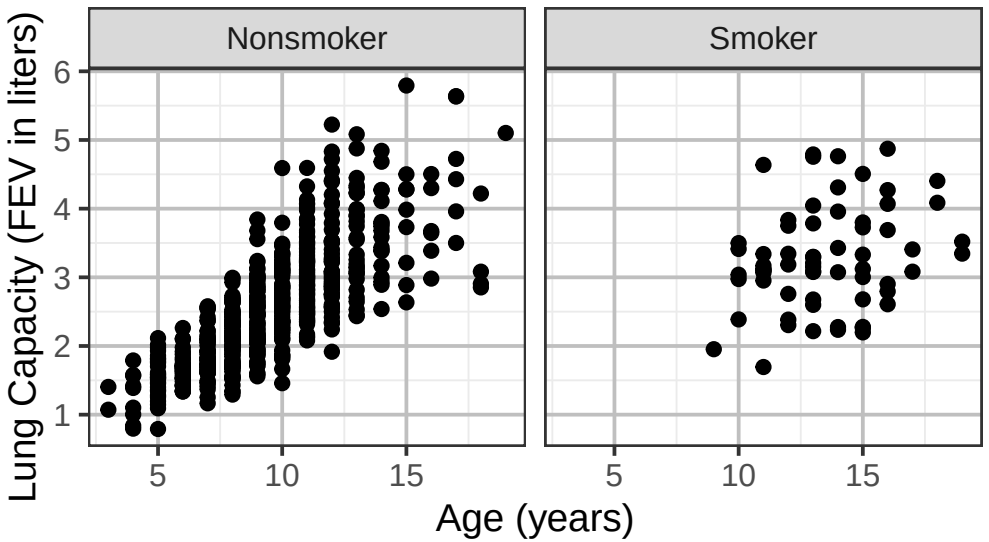


- In `ggplot`, it's easy to make coded scatter plots that the **color** or **shape** of points represent a third variable, like `smoke`
- In FEV data, are smokers or nonsmokers older in general?

Scatter Plots Faceted by Smoke Status

```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_wrap(~smoke) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```

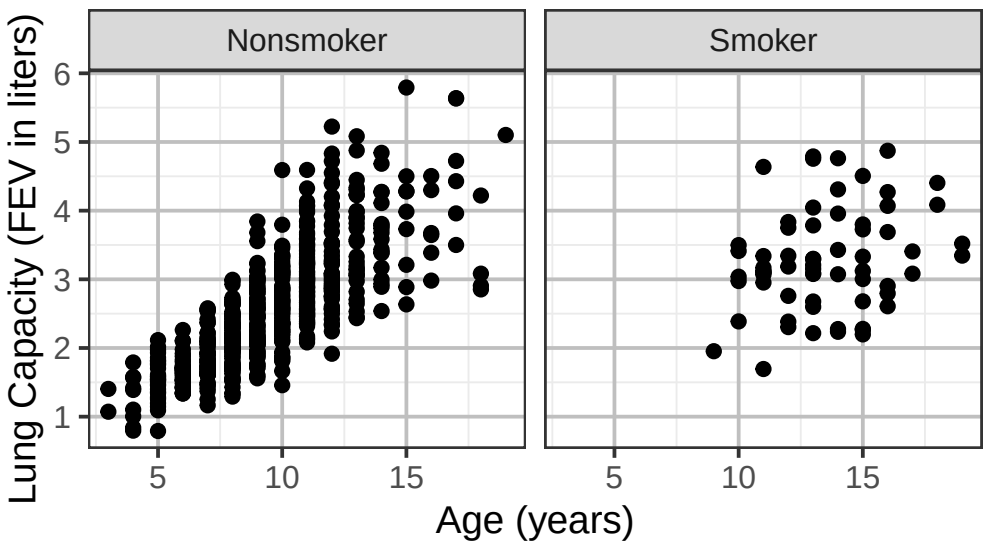
- All subjects are children or teenagers



One can inspect the relation between `age` and `fev` separately for smokers and nonsmokers using `facet_wrap`.

Scatter Plots Faceted by Smoke Status

```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_wrap(~smoke) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```

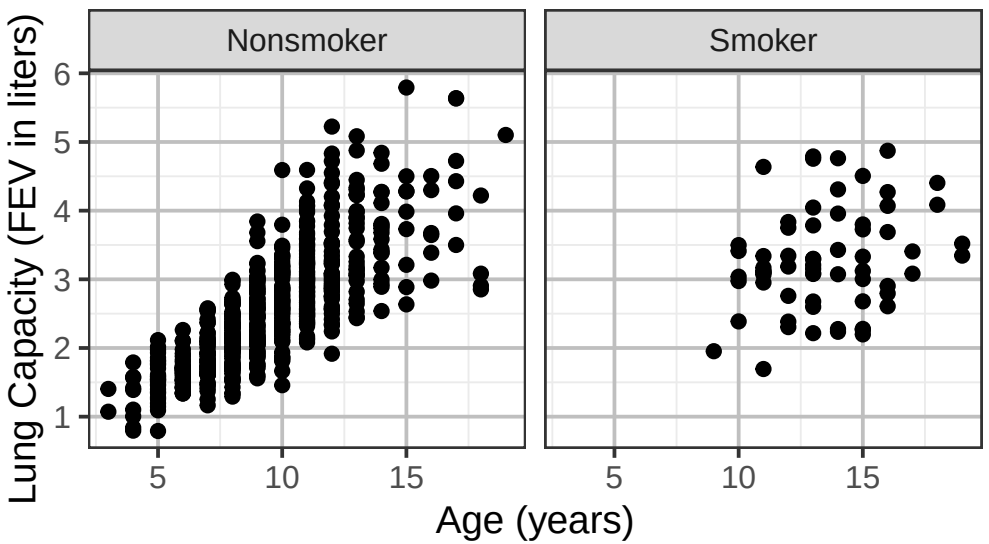


- All subjects are children or teenagers
- Age confounds the relation between smoking status and lung capacities

One can inspect the relation between `age` and `fev` separately for smokers and nonsmokers using `facet_wrap`.

Scatter Plots Faceted by Smoke Status

```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_wrap(~smoke) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```

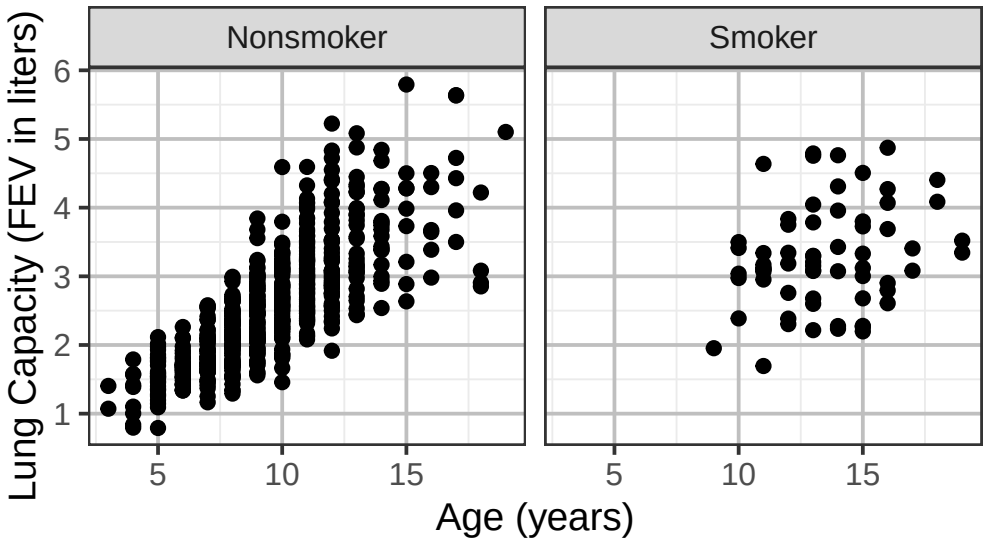


- All subjects are children or teenagers
- Age confounds the relation between smoking status and lung capacities
- Older children had greater lung capacities

One can inspect the relation between `age` and `fev` separately for smokers and nonsmokers using `facet_wrap`.

Scatter Plots Faceted by Smoke Status

```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_wrap(~smoke) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```



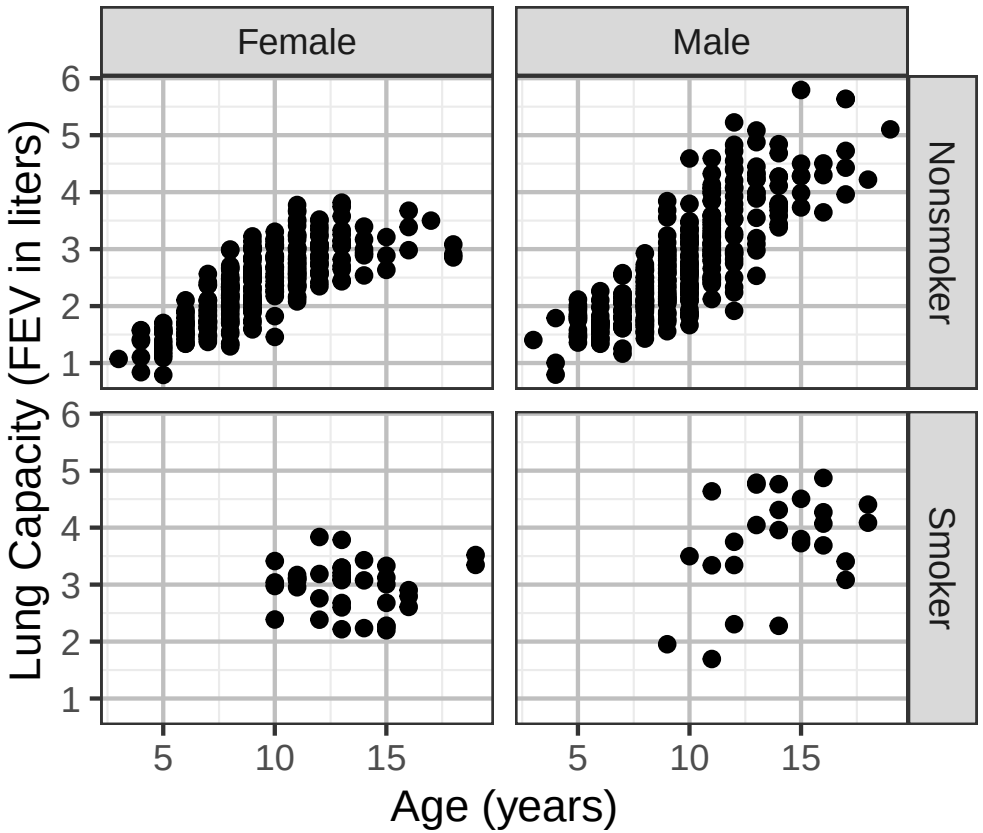
One can inspect the relation between `age` and `fev` separately for smokers and nonsmokers using `facet_wrap`.

- All subjects are children or teenagers
- Age confounds the relation between smoking status and lung capacities
- Older children had greater lung capacities
- Smokers are older (mostly teens) while nonsmokers include children and teens

facet_grid Over Both sex & smoke

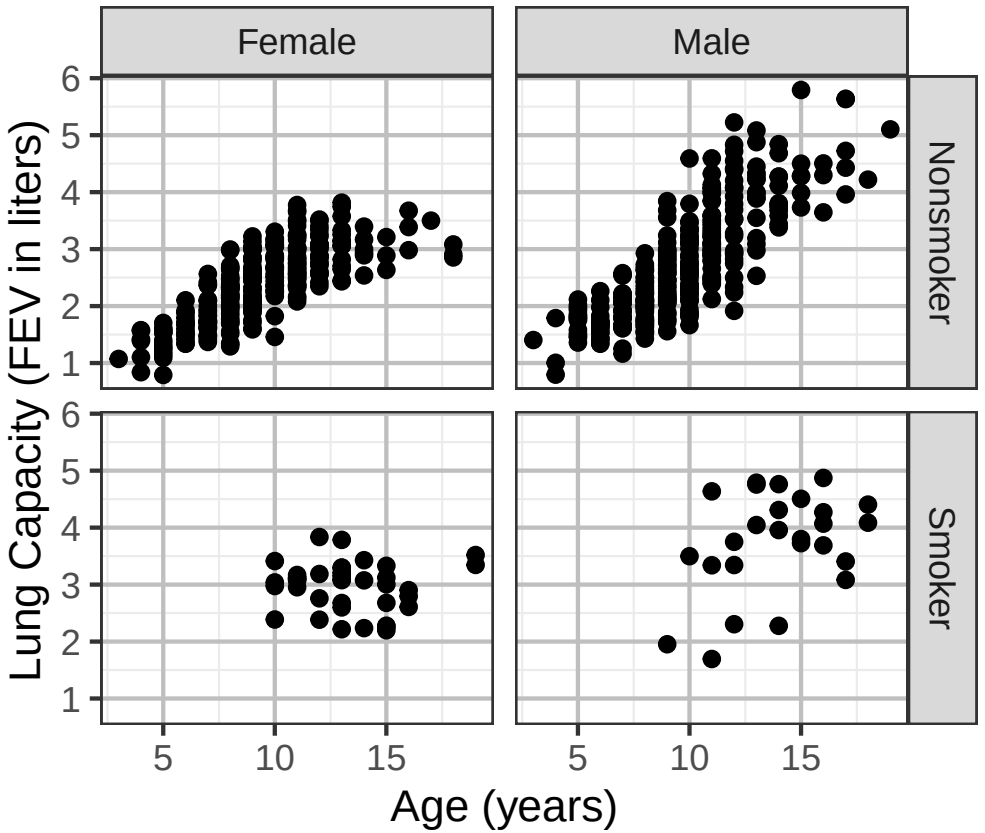
```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_grid(smoke~sex) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```

- Should account for sex since men have greater lung capacities than women.



facet_grid Over Both sex & smoke

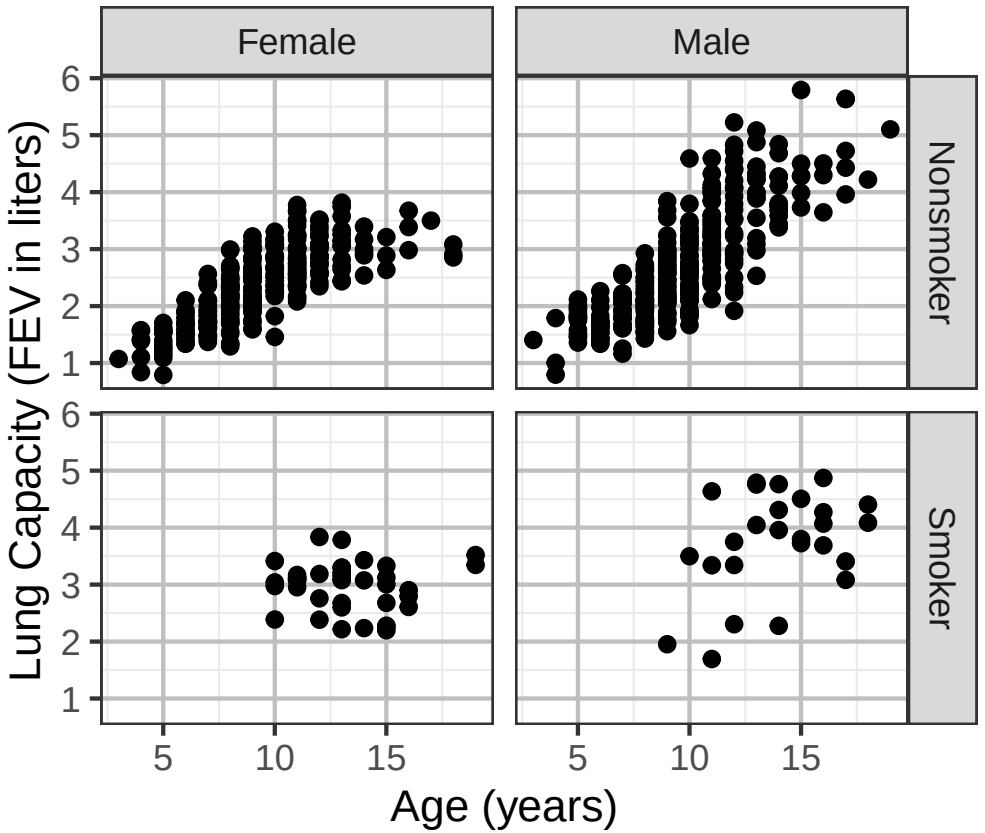
```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_grid(smoke~sex) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```



- Should account for sex since men have greater lung capacities than women.
- For each gender, age still confounds the relation between fev and smoke.

facet_grid Over Both sex & smoke

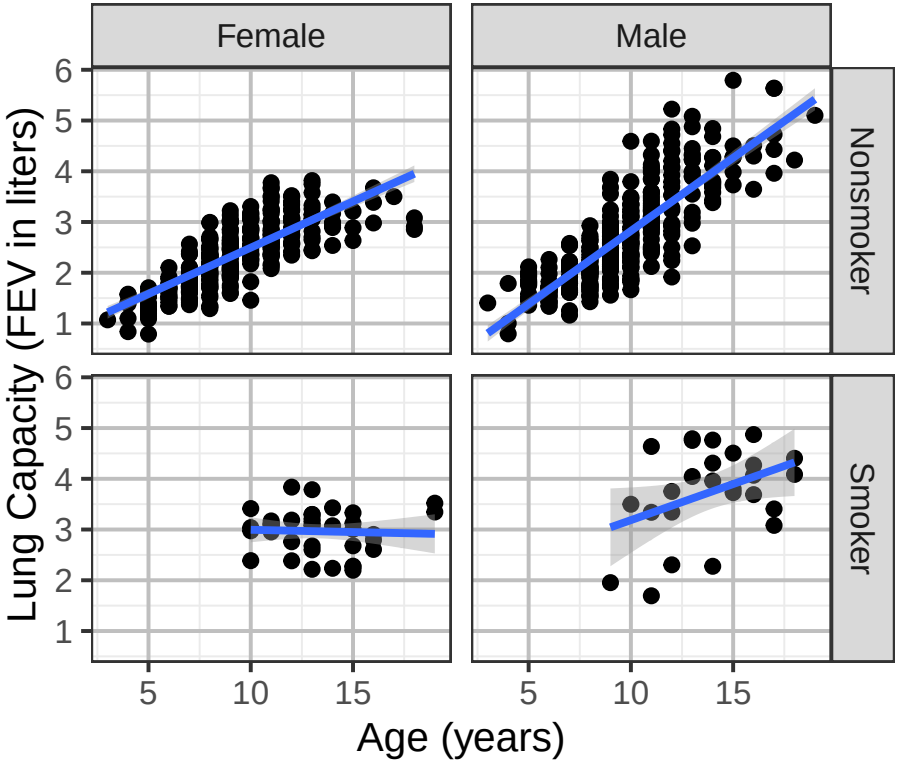
```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_grid(smoke~sex) +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```



- Should account for sex since men have greater lung capacities than women.
- For each gender, age still confounds the relation between fev and smoke.
- Does smoke have any effect on fev after accounting for sex and age?

Adding Least-Square Regression Lines (1)

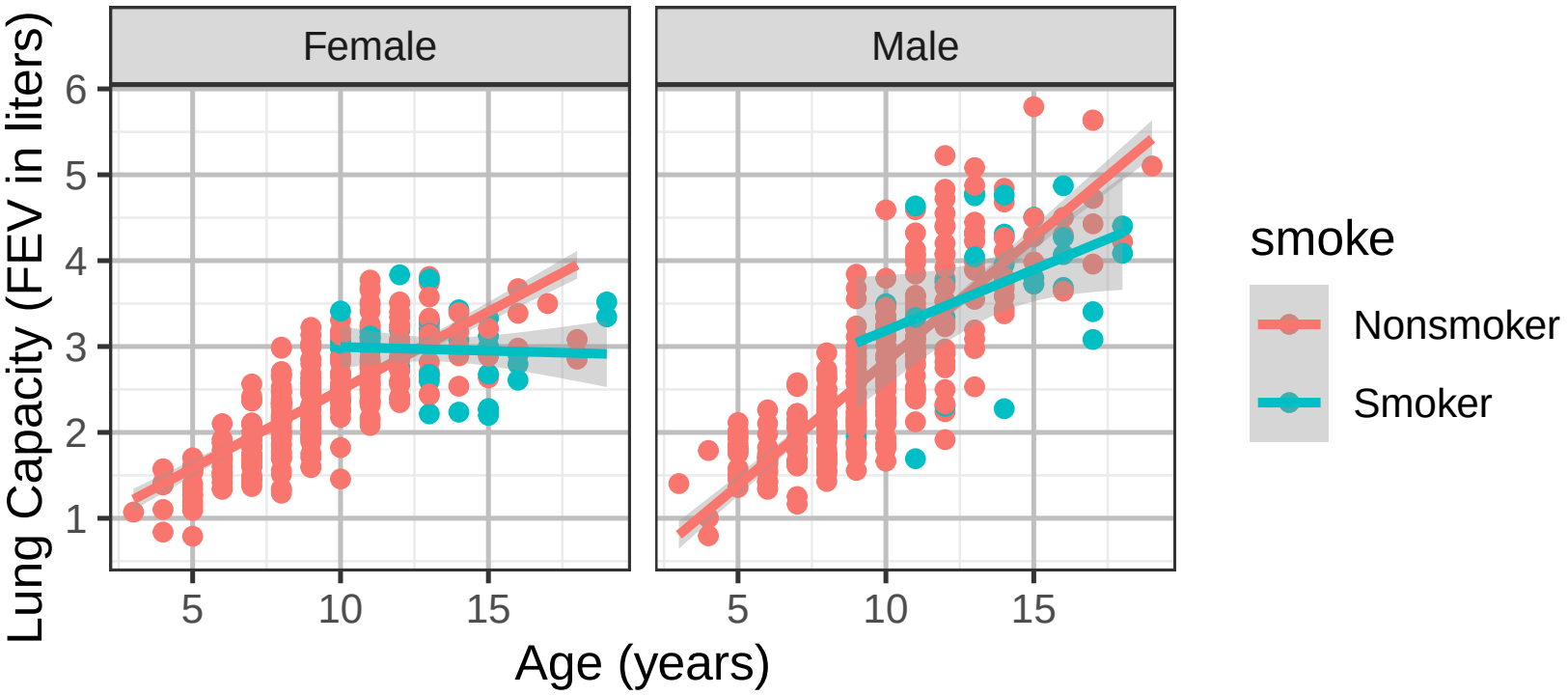
```
ggplot(fevdata, aes(x = age, y = fev)) +
  geom_point() +
  facet_grid(smoke~sex) +
  geom_smooth(method="lm") +
  xlab("Age (years)") +
  ylab("Lung Capacity (FEV in liters)")
```



The LS line for female smokers is clearly less steep than that for female nonsmokers, indicating smoking has a negative effect on the growth of lung capacities for females.

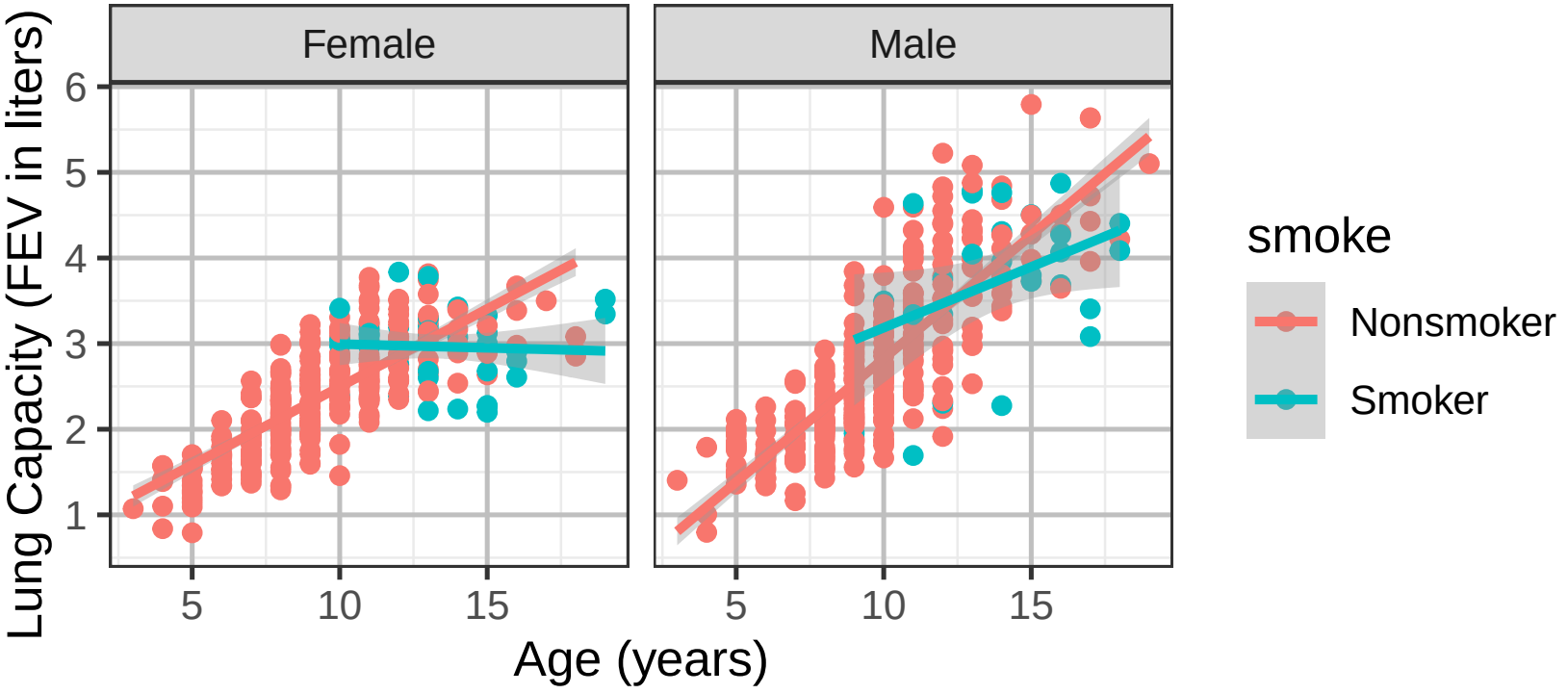
Ditto for males.

```
ggplot(fevdata, aes(x = age, y = fev, col=smoke)) + geom_point() +
  facet_grid(~sex) + geom_smooth(method="lm") +
  xlab("Age (years)") + ylab("Lung Capacity (FEV in liters)")
```



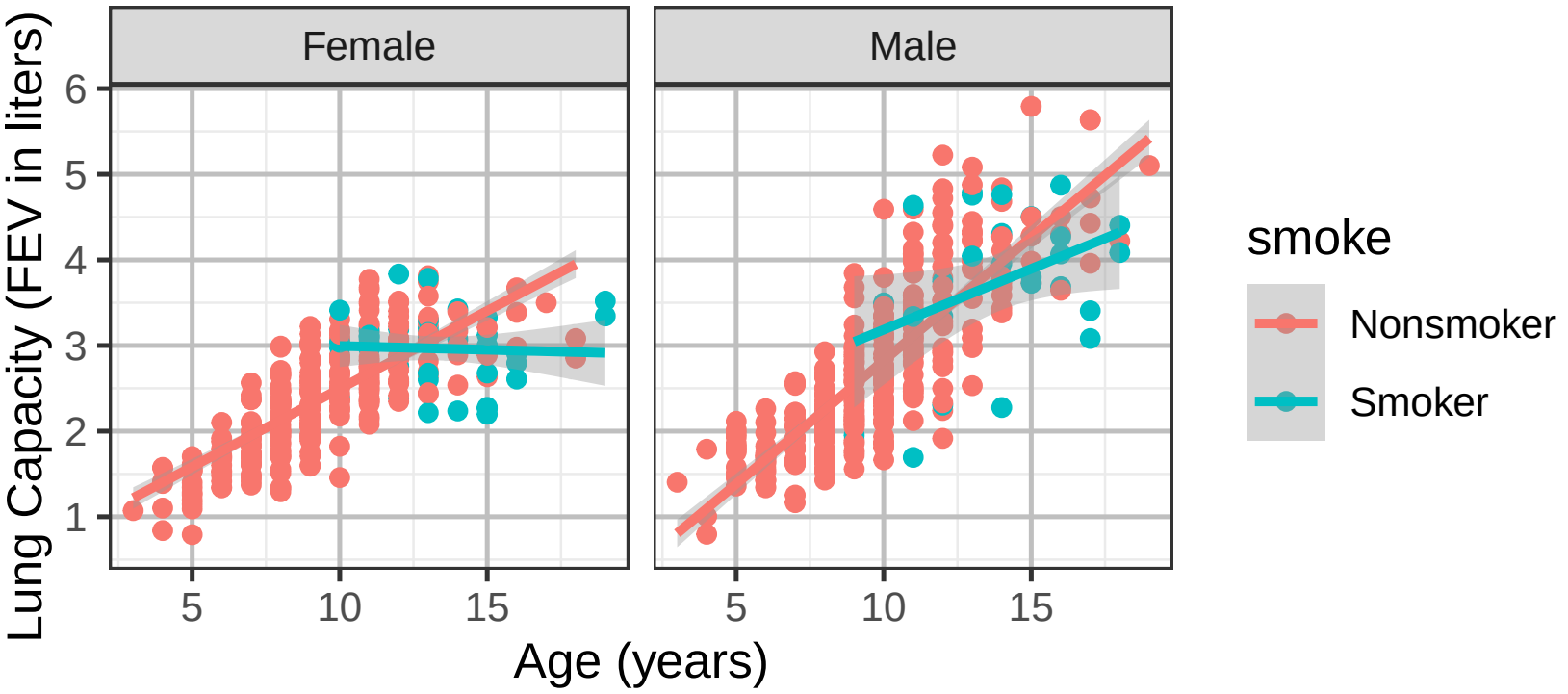
- Easier to compare regression lines when placed on the same plot.

```
ggplot(fevdata, aes(x = age, y = fev, col=smoke)) + geom_point() +
  facet_grid(~sex) + geom_smooth(method="lm") +
  xlab("Age (years)") + ylab("Lung Capacity (FEV in liters)")
```



- Easier to compare regression lines when placed on the same plot.
- Slope for male smokers is also lower than that for male nonsmokers

```
ggplot(fevdata, aes(x = age, y = fev, col=smoke)) + geom_point() +
  facet_grid(~sex) + geom_smooth(method="lm") +
  xlab("Age (years)") + ylab("Lung Capacity (FEV in liters)")
```



- Easier to compare regression lines when placed on the same plot.
- Slope for male smokers is also lower than that for male nonsmokers
- Need a test of whether the slopes of the regression lines are different, which we will do in Chapter 5

A Useful ggplot Tutorial

Introduction to R Graphics with *ggplot2* (from Harvard)

`https://iqss.github.io/dss-workshops/Rgraphics.html`