Lesson 3:
Graphs

Recap: Descriptive Statistics

- Categorical variables: frequency tables, bar plots

- Continuous variables: descriptive measures, histograms

Important commands (recap)

COMMAND	EFFECT
`round(object, x)`	round numbers in object to x decimal places
`par(mfrow = c(x, y), oma = (0,0,0,0))`	environment where graphs will be put in one plot of x rows, y columns
`barplot(var, main = c("title"), ylim = range())`	create barplot of variable, add main title and change range of y-axis
`table(var_1, (var_2))`	create frequency table for one or optionally two variables
`crosstab(var_1, var_2)`	package `descr`, more detailed cross table, including a graph
`mean(var, na.omit = TRUE), median(), sd(), quantile(), range()`	calculate mean/median/sd/quantile/range of variable. Add `na.rm = TRUE` to exclude missings from calculation
`describe(var)`	package `psych`, more extensive than summary() with skew, kurtosis etc.
`describeBy(var, group = cat_var)`	package `psych`, descriptive statistics for variable, separated for group variable
`hist(var, breaks = x, freq = TRUE)`	create histogram of variable, change count of bars, change `freq` to `FALSE` to get density function instead
`curve(dnorm(x, mean = , sd = , ), add = TRUE, col = "red")`	produces a curve that follows a normal distribution with mean and sd, added to last graphic, colour is read

Graphs

Categorical variables

1. bar plot
2. pie chart

Continuous variables

3. histogram
4. boxplot
5. qq plot
6. spaghetti plot

first: short introduction to `ggplot2` package

- `ggplot2` produces very nice plots, but works a little differently than barplot and histogram functions
- ingredients for a plot with ggplot:

1. a dataset (as data frame)
2. aesthetics (x and y axis)
3. shape (geom) layers

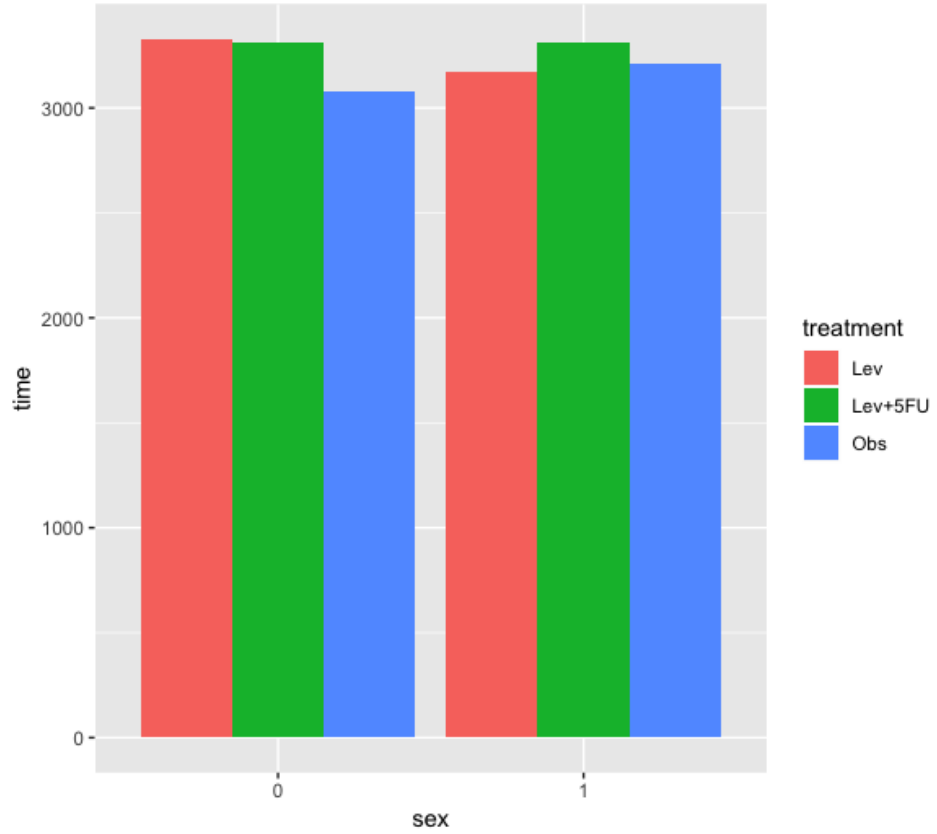
```
ggplot(data = data,
       aes(x = treatment, y = age)) +
  geom_bar(stat = "identity")
```

- easy graphs can also be done without ggplot

1. categorical: barplot

- already know: `barplot()` function
- nicer plot with ggplot2

```
ggplot(data = colon,
       aes(x = sex,
           y = time,
           fill = treatment)) +
  geom_bar(stat = "identity",
           position = position_dodge())
```



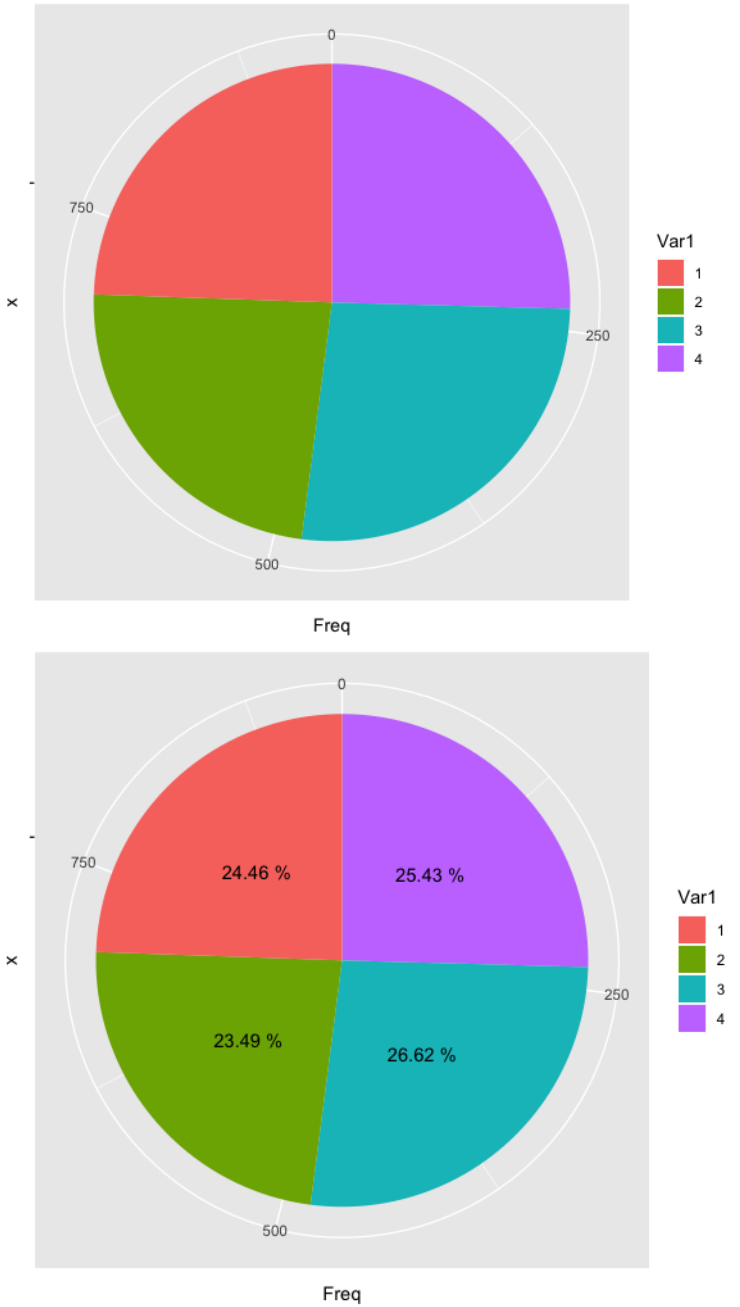
2. categorical: pie chart

- with ggplot

```
ggplot(df, aes(x = "", y = Freq, fill = Var1)) +
  coord_polar("y", start = 0) +
  geom_bar(width = 1, stat = "identity")
```

```
                                                +
  geom_text(aes(label = paste(df$percent, "%")),
            position = position_stack(vjust = 0.5))
```

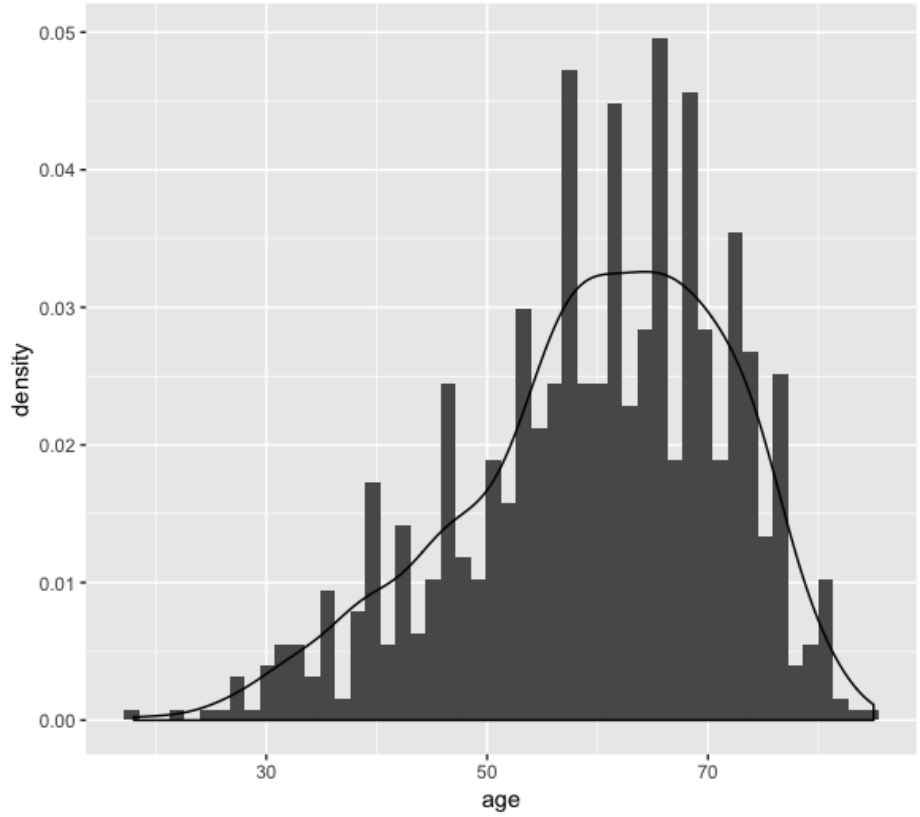
```
                                                +
  ggtitle("follow-up time (death)") +
  theme(plot.title = element_text(hjust = 0.5))
```



3. continuous: histogram

- already know `hist()` function
- nicer plot with ggplot2

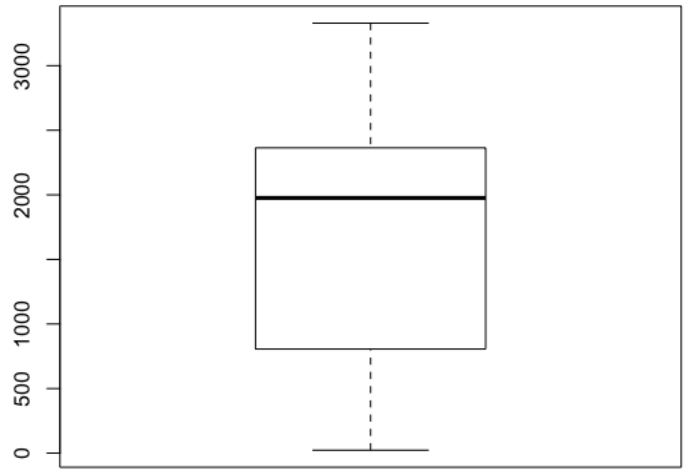
```
ggplot(data = colon, aes(x = age)) +
  geom_histogram(aes(y = ..density..),
                 bins = 50)
  +
  geom_density()
```



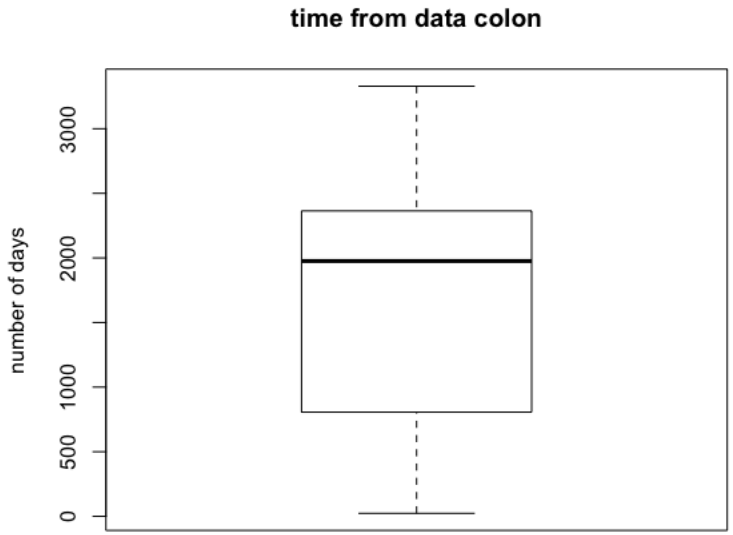
4. continuous: boxplot

- without ggplot2:

```
boxplot(colon$time)
```



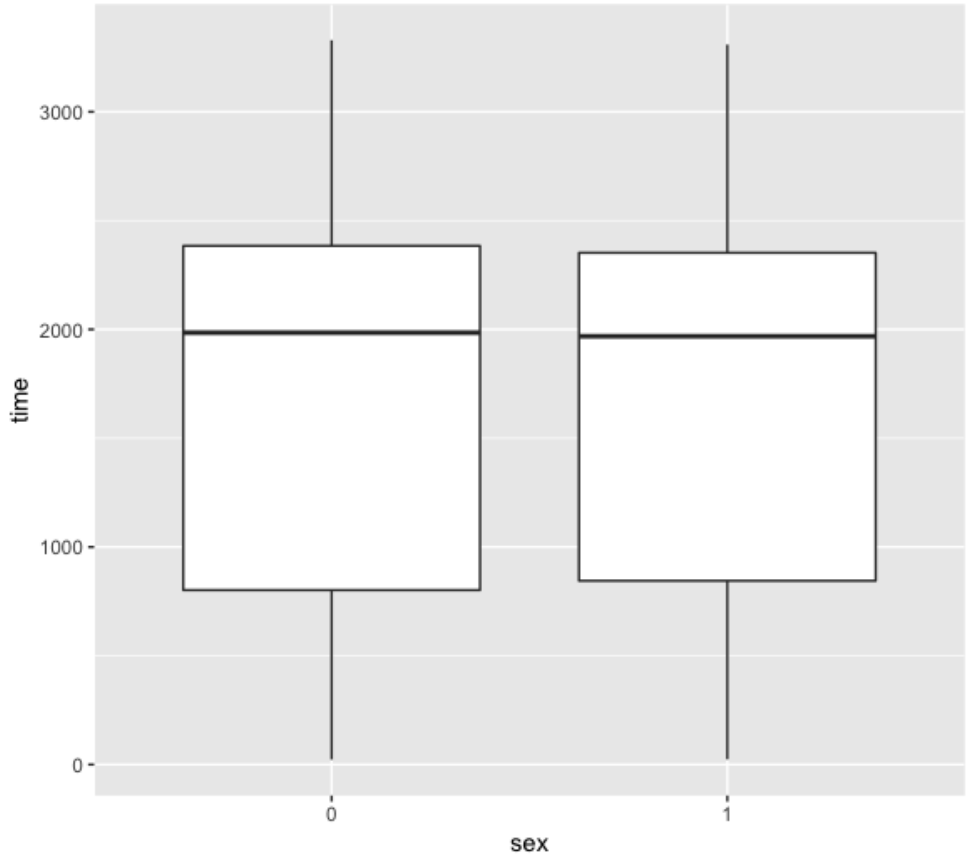

```
boxplot(colon$time,
        main = "time from data colon",
        ylab = "number of days",
        range = max(colon$time,
                    na.rm = TRUE))
```



4. continuous: boxplot

- with ggplot2

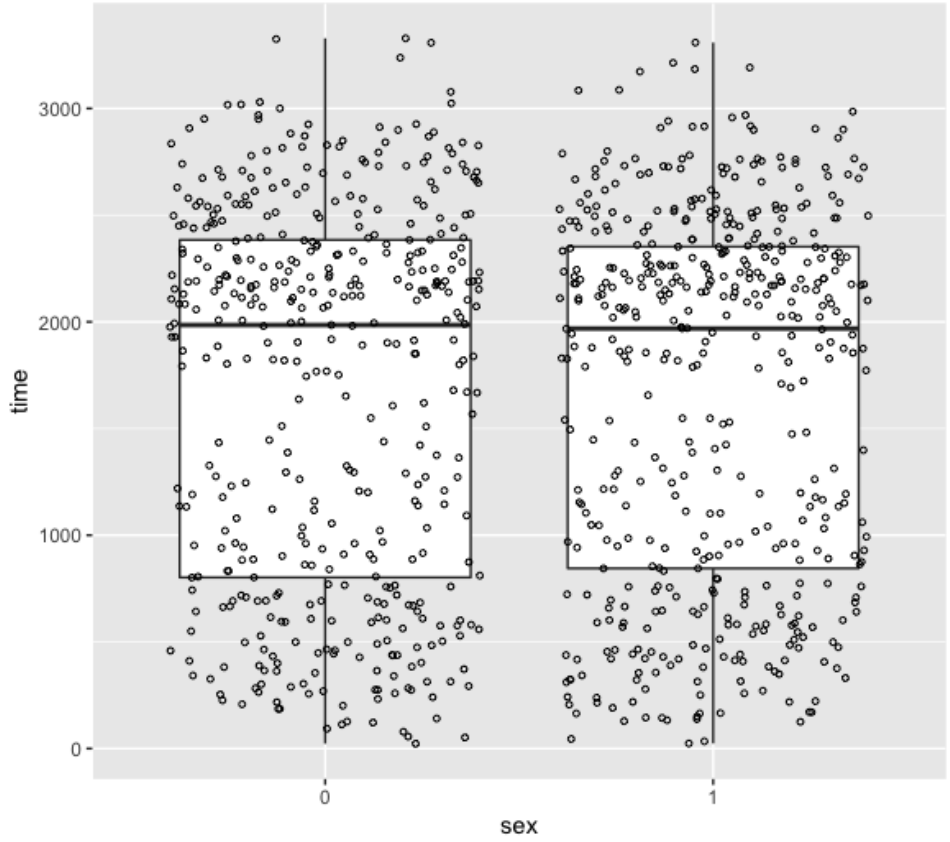
```
ggplot(data = colon, aes(x = sex, y = time)) +
  geom_boxplot()
```



4. continuous: boxplot

- with ggplot2

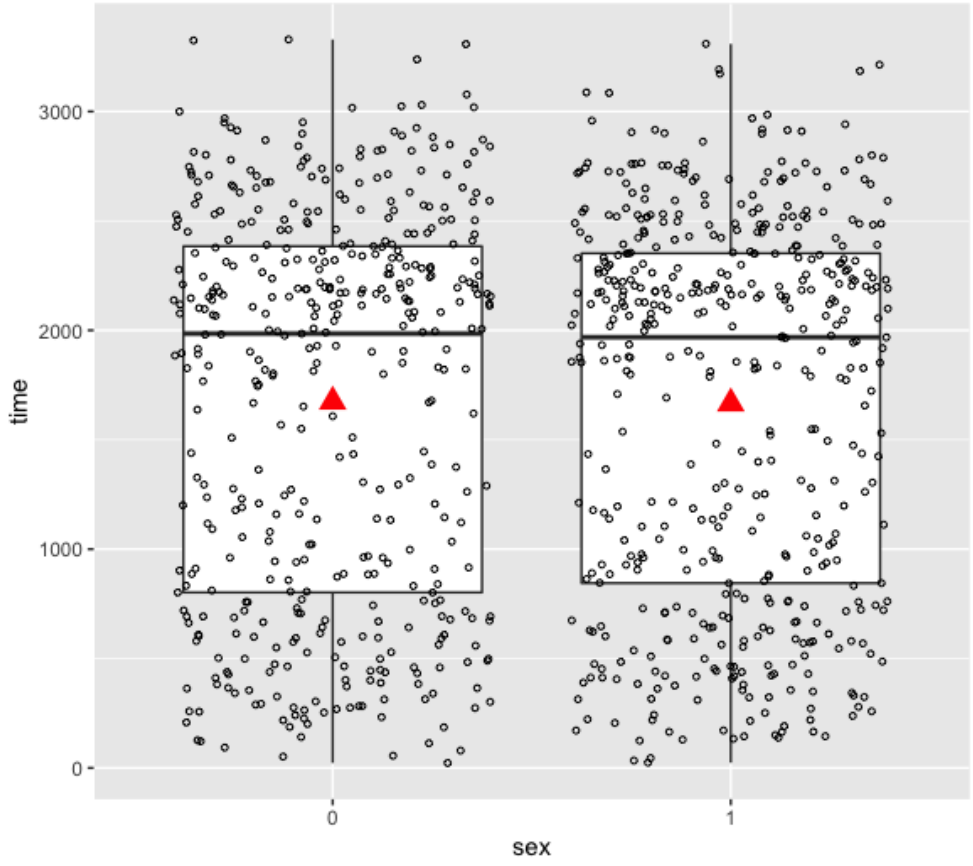
```
ggplot(data = colon, aes(x = sex, y = time)) +
  geom_boxplot() +
  geom_jitter(colour = "black",
                  size = 1,
                  shape = 1)
```



4. continuous: boxplot

- with ggplot2

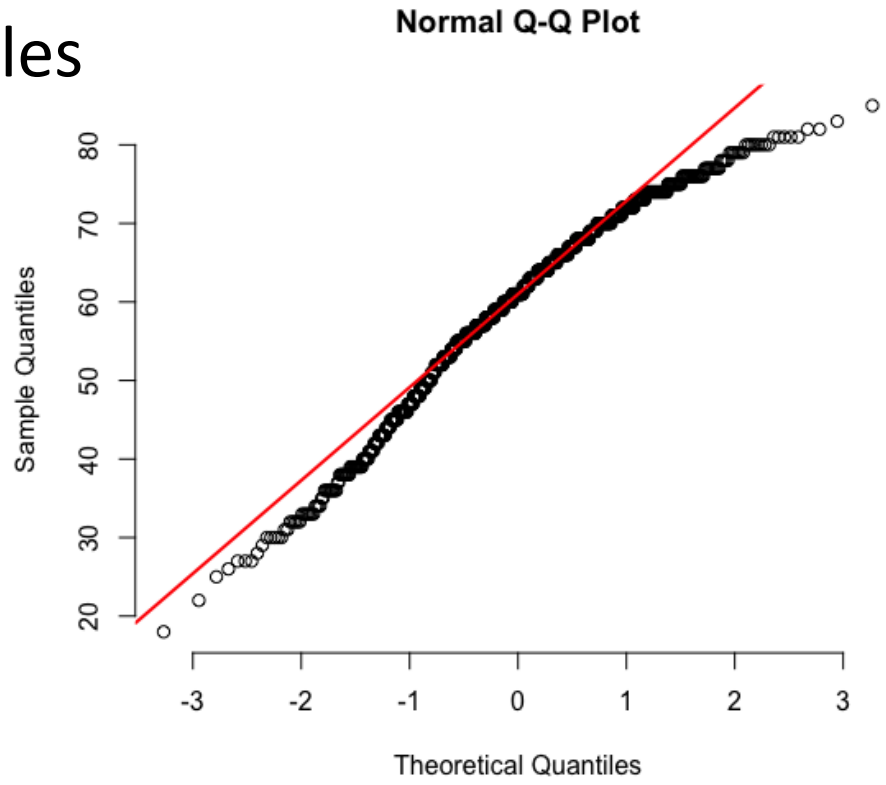
```
ggplot(data = colon, aes(x = sex, y = time)) +
  geom_boxplot() +
  geom_jitter(colour = "black",
              size = 1,
              shape = 1) +
  stat_summary(fun.y = mean,
               geom = "point",
               shape = 17,
               size = 4,
               col = "red")
```



5. continuous: qq plot

- to determine if normally distributed
- compare data to normal distributed quantiles

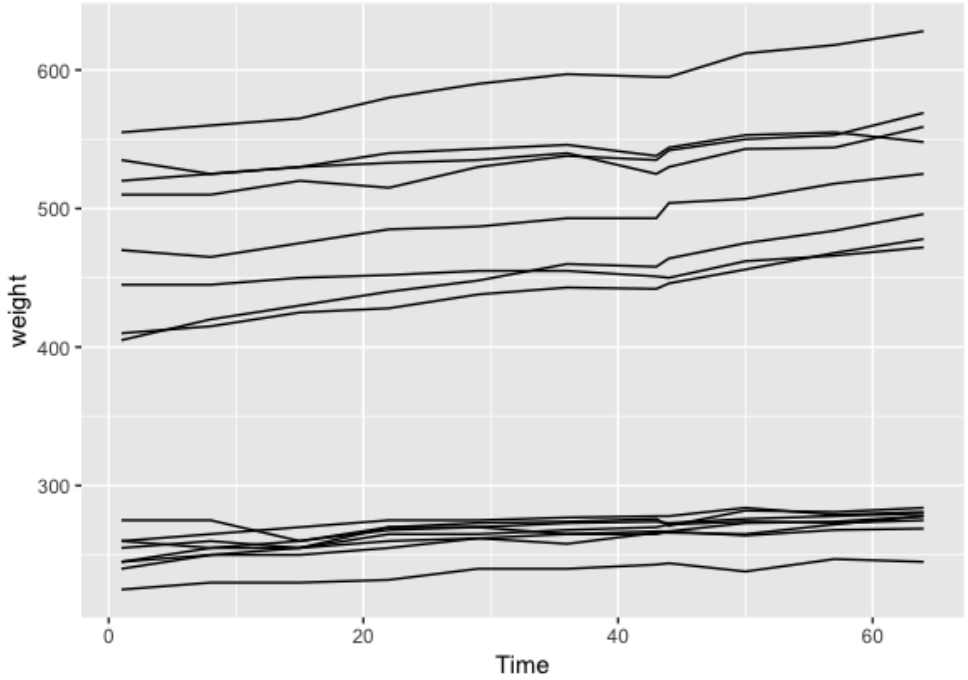
```
qqnorm(colon$age, pch = 1, frame = FALSE)
qqline(colon$age, col = "red", lwd = 2)
```



6. continuous: spaghetti plot

- for longitudinal data

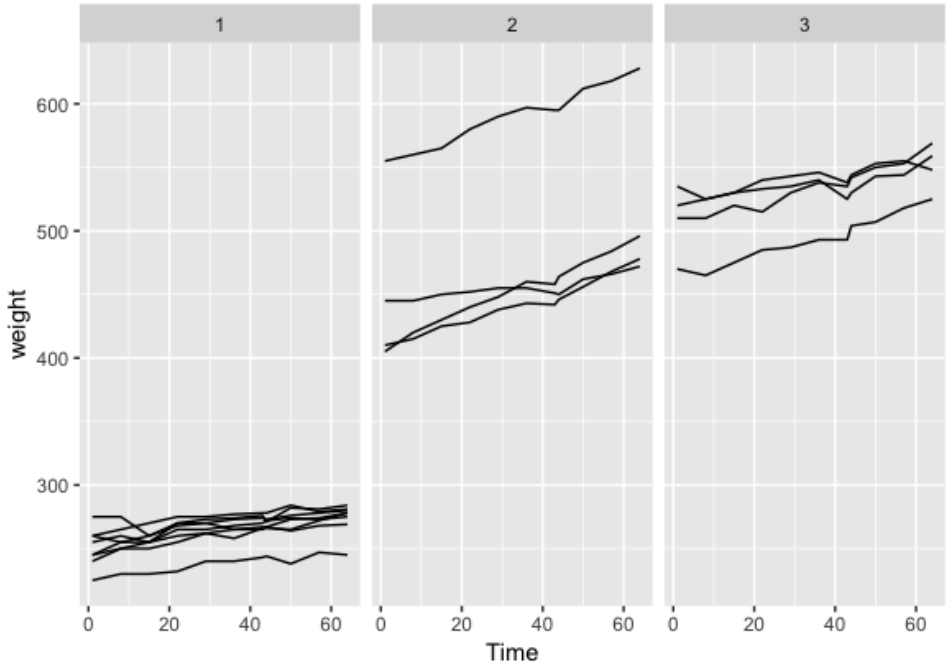
```
ggplot(data = rats_weight,
       aes(x = Time, y = weight, group = Rat)) +
  geom_line()
```



6. continuous: spaghetti plot

- for longitudinal data

```
ggplot(data = rats_weight,
       aes(x = Time, y = weight, group = Rat)) +
  geom_line() +
  facet_grid(. ~ Diet)
```



Important commands

COMMAND	EFFECT
`boxplot(var)`	boxplot of a variable (without using `ggplot2`)
`ggplot(data = data, aes(x = var1, y = var2, fill = var_cat))`	package `ggplot2`, needs to be specified with geom layers (see below)
`coord_polar(start = 0, "y")`	make pie chart from data
`geom_bar(stat = "identity", position = position_dodge(), width = 1)`	bar plot (with bars arranged next to each other)
`geom_text()`	add text (such as a label) to the plot
`ggtitle()`	add a title to the plot
`theme()`	
`geom_histogram(aes(y = y), bins = 10)`	histogram with specified number of bins, `y = .. density ..` for density plot
`geom_density()`	add density curve to the histogram plot
`geom_boxplot()`	boxplot of the data
`geom_jitter()`	add data points to the boxplot

Important commands (continued)

COMMAND	EFFECT
`stat_summary()`	add statistics to the boxplot, e.g. mean of data
`geom_line()`	data as line in plot
`facet_grid()`	split data up into subgroups
`qqnorm(var, pch = 1, frame = FALSE)`	data as a qq plot with quantiles
`qqline(var, col = "red", lwd = 1)`	adds quantile line of normal distribution