STAD29 / STA 1007 assignment 6

Due Tuesday March 10 at 11:59pm on Quercus

Packages for this one:

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Hand in problems 2 and 4.

1. Work through the rest of Chapter 22 of PASIAS.

2. Back in STAC32, we had some children who were learning to read. We now have some more, but the experimenters were concerned that the total income of each child's family might also affect the child's reading score. (You might imagine that a larger family income, other things being equal, would be associated with a higher (better) reading score.) There are, this time, four reading methods, labelled `method1` through `method4`. The data for this study are in `http://ritsokiguess.site/STAD29/reading_again.csv`.

   (a) (2 marks) Read in the data and display (some of) the data frame.

Solution: Exactly the usual:

```
my_url <- "http://ritsokiguess.site/STAD29/reading_again.csv"
scores <- read_csv(my_url)

## Parsed with column specification:
## cols(
##   method = col_character(),
##   reading = col_double(),
##   income = col_double()
## )

scores

## # A tibble: 36 x 3
##    method  reading income
##    <chr>     <dbl>  <dbl>
##  1 method1      12   17.5
##  2 method2      45   70.8
##  3 method3      20   71.4
##  4 method4      12   35
##  5 method1      39  105.
## 6 method2      37   45.9
## 7 method3      42   55
## 8 method4      10   33
## 9 method1      36   64.7
## 10 method2     13   47.5
## # ... with 26 more rows
```

You note that there are columns of method for learning to read (with the right labels), a reading score, and a family income value (evidently in thousands of dollars).

Extra: the data came from `http://www.real-statistics.com/analysis-of-covariance-ancova/basic-concepts-ancova/`, where you see the values are actually in two separate tables, and the values of reading score and family income in the corresponding places in each table are for the same child. To get these data into the form that you have, I first duplicated the tables on the website:

```
reading1 <- tribble(
  ~method1, ~method2, ~method3, ~method4,
  12, 45, 20, 12,
  39, 37, 42, 10,
  36, 13, 31, 19,
  17, 50, 24, 18,
  25, 35, 15, 14,
  15, 40, 13, 8,
  8, 33, 9, 7,
  31, 17, 21, 19,
  NA, NA, 31, 25,
  NA, NA, 13, 26
)
reading1
## # A tibble: 10 x 4
##    method1 method2 method3 method4
##      <dbl>   <dbl>   <dbl>   <dbl>
##  1      12      45      20      12
##  2      39      37      42      10
##  3      36      13      31      19
##  4      17      50      24      18
##  5      25      35      15      14
##  6      15      40      13       8
##  7       8      33       9       7
##  8      31      17      21      19
##  9      NA      NA      31      25
## 10      NA      NA      13      26
income <- tribble(
  ~method1, ~method2, ~method3, ~method4,
  17.5, 70.8, 71.4, 35,
  104.6, 45.9, 55, 33,
  64.7, 47.5, 54, 34.2,
  47, 77.8, 27.9, 43.2,
  22, 70.9, 40.6, 20,
  12.4, 84.8, 33, 37,
  20, 49.8, 22.2, 28.2,
  79.7, 34.6, 80.5, 46.4,
  NA, NA, 80, 64.9,
  NA, NA, 41, 59.4
)
income
```

```
## # A tibble: 10 x 4
##    method1 method2 method3 method4
##      <dbl>   <dbl>   <dbl>   <dbl>
##  1    17.5    70.8    71.4    35
##  2   105.     45.9    55      33
##  3    64.7    47.5    54      34.2
##  4    47      77.8    27.9    43.2
##  5    22      70.9    40.6    20
##  6    12.4    84.8    33      37
##  7    20      49.8    22.2    28.2
##  8    79.7    34.6    80.5    46.4
##  9    NA      NA      80      64.9
## 10    NA      NA      41      59.4
```

There were two fewer children doing methods 1 and 2, hence the missings in the tables.

The first job is to pivot-longer each of these:

```
reading1 %>% pivot_longer(everything(), names_to="method", values_to="reading") -> d1
d1
```

```
## # A tibble: 40 x 2
##    method  reading
##    <chr>     <dbl>
##  1 method1      12
##  2 method2      45
##  3 method3      20
##  4 method4      12
##  5 method1      39
##  6 method2      37
##  7 method3      42
##  8 method4      10
##  9 method1      36
## 10 method2      13
## # ... with 30 more rows
```

```
income %>% pivot_longer(everything(), names_to="method", values_to="income") -> d2
d2
```

```
## # A tibble: 40 x 2
##    method  income
##    <chr>    <dbl>
##  1 method1   17.5
##  2 method2   70.8
##  3 method3   71.4
##  4 method4   35
##  5 method1  105.
##  6 method2   45.9
##  7 method3   55
##  8 method4   33
##  9 method1   64.7
## 10 method2   47.5
## # ... with 30 more rows
```

Realizing that a row in each table goes with the corresponding row in the other table, we can glue them together using `bind_cols`:

```
d1 %>%  bind_cols(d2) %>% select(-method1) %>% drop_na() -> reading_again
reading_again
```

```
## # A tibble: 36 x 3
##    method  reading income
##    <chr>     <dbl>  <dbl>
##  1 method1      12   17.5
##  2 method2      45   70.8
##  3 method3      20   71.4
##  4 method4      12   35
##  5 method1      39  105.
##  6 method2      37   45.9
##  7 method3      42   55
##  8 method4      10   33
##  9 method1      36   64.7
## 10 method2      13   47.5
## # ... with 26 more rows
```

`method` appears in both tables, so we can get rid of the duplicate (which acquired the name `method1`), and also get rid of the rows with missing values.

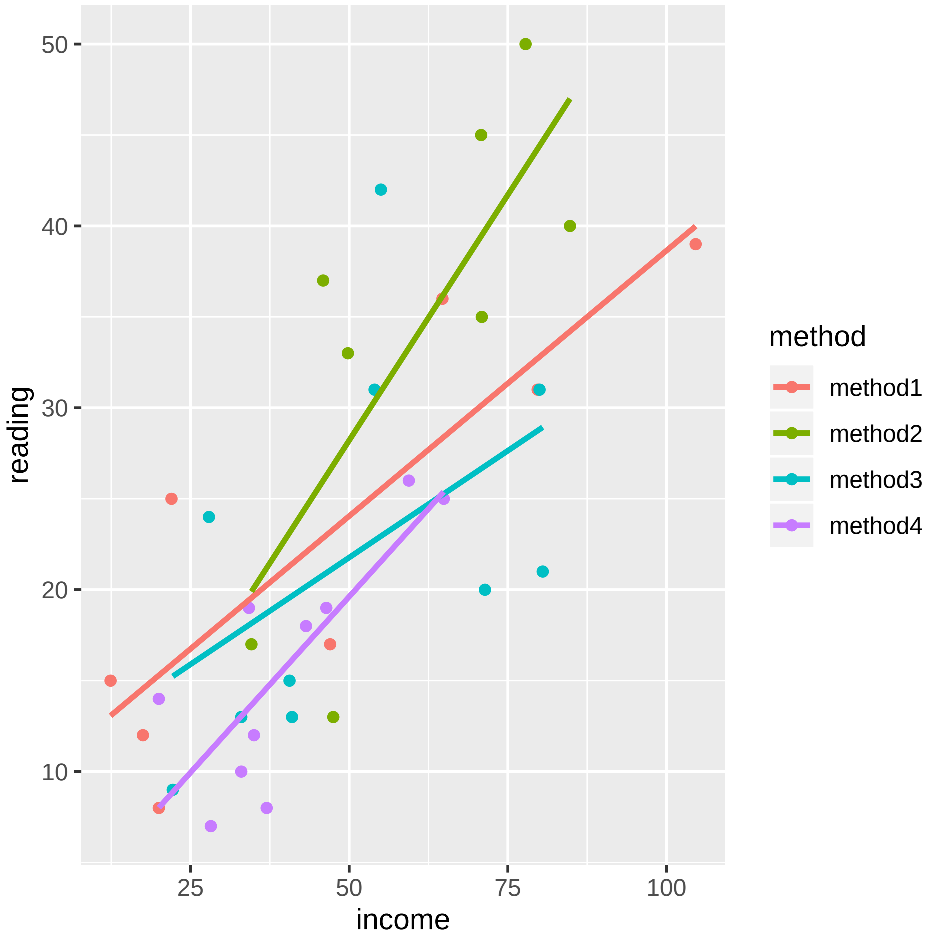
Oftentimes we combine two tables with something like `left_join`. But that won't work here since we have nothing like a child ID that we can match on. We had to be lucky enough to have the rows match up so that we could use `bind_cols`.

Finally, I used `write_csv` to write this to a file, which is the one you read in.

(b) (3 marks) Make a suitable plot of the data. Add regression lines for each method (*without* the grey envelopes). Bear in mind that we are trying to predict reading score from everything else.

Solution: The usual plot for these: a scatterplot of the two quantitative variables, with the groups distinguished by colour (reading methods). The reading score is the response, so it goes on the y axis:

```
ggplot(scores, aes(x=income, y=reading, colour=method)) + geom_point() +
    geom_smooth(method="lm", se=F)
```



You need both those things in the `geom_smooth`, to get regression lines rather than smooth trends, and to *not* get the grey envelopes around the regression lines. (If you have them here, they overlap rather and obscure things, so I think it's better to not have them.)

(c) (2 marks) Describe any effects of income and reading method on reading score that you see on the graph.

Solution: The effect of income is that reading score tends to increase as income increases (the lines all go uphill). The effect of reading method seems to be that method 2 is the best for most incomes (its line is the highest for any income that was observed in the data) and that method 4 is the worst (its line is the lowest for any observed income). Say something sensible here; I don't mind so much what as long as you say *something*.

(d) (2 marks) Run an analysis of variance of reading score as it depends on reading method. Display the results.

Solution: This is from C32:

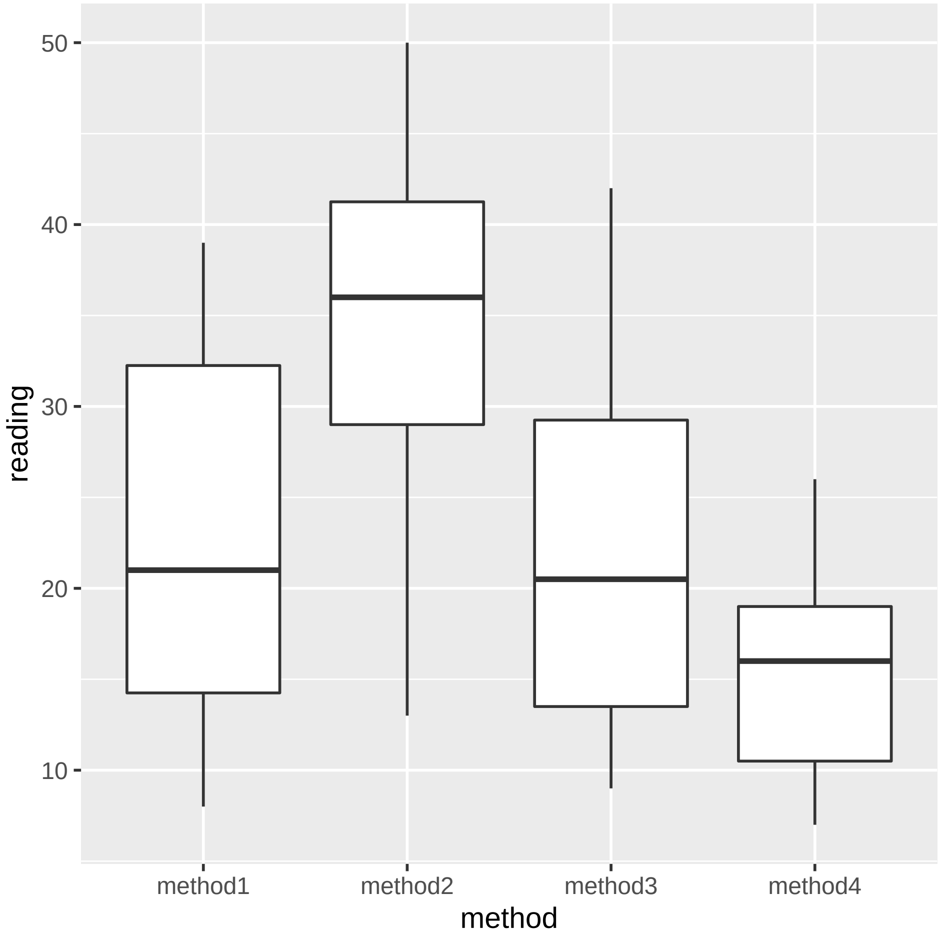
```
scores.1 <- aov(reading~method, data=scores)
summary(scores.1)
##             Df Sum Sq Mean Sq F value  Pr(>F)
## method       3   1455   485.0   4.503 0.00958 **
## Residuals   32   3447   107.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I don't need Tukey, since the aim here is to compare this ANOVA with an analysis of covariance (and think about which one is more suitable). Note the P-value here is small, so that according to this the mean reading score is not the same for each reading method.

(e) (3 marks) Compare, using a suitable graph or numerical summary, the reading scores for the different reading methods. What is your main conclusion?

Solution: One of these two:

```
ggplot(scores, aes(x=method, y=reading)) + geom_boxplot()
```



or

```
scores %>% group_by(method) %>% summarize(m=mean(reading))
```

```
## # A tibble: 4 x 2
##   method      m
##   <chr>   <dbl>
## 1 method1  22.9
## 2 method2  33.8
## 3 method3  21.9
## 4 method4  15.8
```

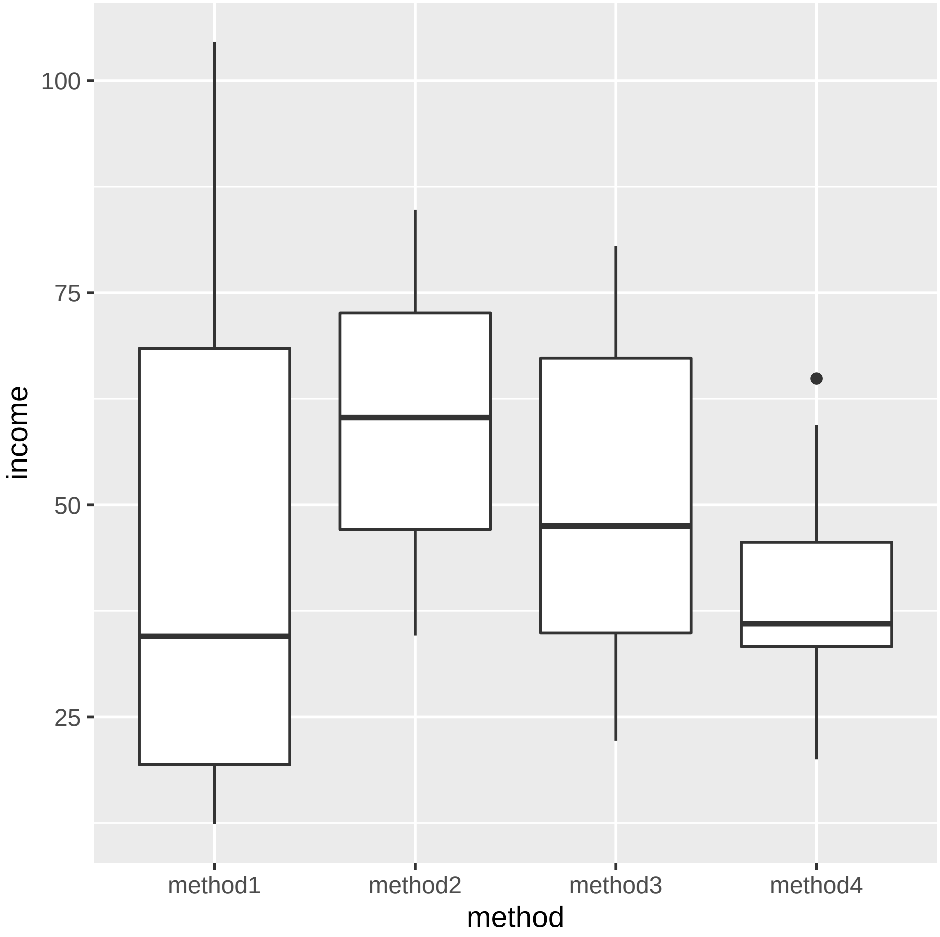
(or, if you like, the median reading score instead of the mean).

I think the main conclusion ought to be that method 2 is better than the others (which are not very different).

(f) (3 marks) Repeat the previous part, but this time comparing the family income by reading method (and not the reading scores). Again, comment briefly.

Solution: Same choices:

```
ggplot(scores, aes(x=method, y=income)) + geom_boxplot()
```



or

```
scores %>% group_by(method) %>% summarize(m=mean(income))
```

```
## # A tibble: 4 x 2
##   method      m
##   <chr>   <dbl>
## 1 method1  46.0
## 2 method2  60.3
## 3 method3  50.6
## 4 method4  40.1
```

Method 2 is also associated with the highest family income.

You might have decided to run an ANOVA here, which actually shows that there is no significant difference in income between methods. This obscures the point I make below. My intention was that you would do something that compares the incomes by method in an exploratory way, rather than doing anything inferential. But you see this kind of approach in papers all the time: some kind of comparison that says the treatment groups do not differ on other variables. The implication is that the methods do not differ by income, and therefore that method 2 looks best because it really is best, and not just because it is propped up by high-income students. If you want to go that way, I am fine with it as an answer (to (h)), although I don't really agree with it. My take (see below) is that the groups might differ by income, and we have a "free" way to adjust for that (the ANCOVA) that we might as well take advantage of.

Extra: my graph and summaries ought to make you suspicious that the superiority of method 2 really should not be as big as it appears, because those children also tend to come from families with higher income, and higher family income is also associated with higher reading scores.

(g) (3 marks) Run a suitable analysis of covariance, and use `drop1` with `test="F"` to test the significance of the two explanatory variables. What do you conclude?

Solution: This is a regression predicting reading score from reading method and income (in either order), thus:

```
reading.2=lm(reading~method+income, data=scores)
drop1(reading.2, test="F")
```

```
## Single term deletions
##
## Model:
## reading ~ method + income
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              1768.5 150.20
## method  3    571.03 2339.6 154.27  3.3365   0.03194 *
## income  1   1678.35 3446.9 172.22 29.4194 6.367e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both P-values are smaller than 0.05, so reading score depends on both reading method and family income (neither can be removed from the model).

(h) (3 marks) Compare the P-values for `method` from the analysis of covariance in the previous part, and the analysis of variance you did earlier. Which one do you think is more trustworthy? Explain briefly.

Solution: In the ANCOVA in the previous part, the P-value is 0.032; in the earlier ANOVA it was 0.0096. I trust the ANCOVA more, because it properly accounts for the effect of family income, which we know to be important. The ANOVA counts all the variation due to income as "error", which is not appropriate when we know that income has an effect.

"The ANOVA because its P-value is smaller" is clearly *wrong*, because how do you know you can trust that smaller P-value?

The reason for the difference was discussed in the Extra to the previous part: method 2 has the highest reading scores, but also the highest family income, so that we don't know (without extra thought) *why* it is that the reading scores for method 2 are the highest: it could be because method 2 is really best, or it could be because these are mostly children from high-income families, or a mixture of both. We actually know from the ANCOVA that it is a mixture of both, since both explanatory variables have an effect; the story is that method 2 is best, but not by as much as you might think. If you go back to that scatterplot you made (much) earlier, method 2 is best for most incomes, but the amount by which it is best (if it is) depends on income, and is not really as much as you might have guessed from the boxplot of reading score and reading method.

If you did an ANOVA for income and method above and found no significant difference, I suppose you could make an argument here that the method groups do not significantly differ by income, and therefore comparing the method means without considering income is reasonable (and therefore the ANOVA for reading scores by method (only) is trustworthy). It's an argument you see (and therefore I would accept it here if you make it clearly), but I don't really like it (for reasons discussed above)[1]. I think that reading scores depend on both method and income, and we do best to model reading scores in terms of both, to get the most accurate picture.

3. Work through Chapter 23 of PASIAS.

4. Researchers are comparing different ways to give technical information about diet. Specifically, 33 subjects are randomly assigned to one of three groups. The first group receives technical dietary information from a website. Group 2 receives the same information from a nurse practitioner, while group 3 receives the information from a video made by the same nurse practitioner. Each subject then made three ratings: difficulty, usefulness, and importance of the information in the presentation. The researcher looks at the three different ratings of the presentation to determine if there is a difference between the modes of presentation. In particular, the researcher is interested in whether the website is superior because that is the most cost-effective way of delivering the information. In the dataset, the ratings are presented in the variables useful, difficulty and importance. The variable `group` indicates the group to which a subject was assigned.

   The data are in `http://ritsokiguess.site/STAD29/dietary.csv`.

   (a) (2 marks) Read in and display (some of) the data. Make some kind of comment about whether you have what you were expecting.

Solution: All very familiar:

```
my_url <- "http://ritsokiguess.site/STAD29/dietary.csv"
dietary <- read_csv(my_url)
```

```
## Parsed with column specification:
## cols(
##   group = col_character(),
##   useful = col_double(),
##   difficulty = col_double(),
##   importance = col_double()
## )
```

```
dietary
```

```
## # A tibble: 33 x 4
##     group   useful difficulty importance
##     <chr>    <dbl>      <dbl>      <dbl>
##   1 website   19.6       5.15       9.5
##   2 website   15.4       5.75       9.10
##   3 website   22.3       4.35       3.30
##   4 website   24.3       7.55       5
##   5 website   22.5       8.5        6
##   6 website   20.5      10.2        5
##  7 website   14.1       5.95      18.8
##  8 website   13         6.30      16.5
##  9 website   14.1       5.45       8.90
## 10 website   16.7       3.75       6
## # ... with 23 more rows
```

I have 33 rows (one per person); each person is identified by the group they were in, as well as by the three ratings they gave.

(b) (2 marks) Explain briefly why MANOVA is something we would consider to analyze these data.

Solution: We have three quantitative variables that are all outcomes (that is to say, all response variables), and they each (might) depend on one categorical explanatory variable `group`. This suggests a one-way multivariate analysis of variance as a way to analyze the data.

Key: note more than one response variable and one (or more, in general) categorical explanatory variable.

Extra: the value of doing a MANOVA here might be like the seed yield and seed weight example in class, in that it might be a *combination* of the response variables that distinguishes the groups. That is something three ordinary ANOVAs, one for each response variable, could miss.

(c) (4 marks) Make side-by-side boxplots of each of the three explanatory variables for each of the three groups. For full credit, do this in *one* `ggplot`. Hint: use the idea from C32 for plotting residuals against each of the explanatory variables in a regression, all in one plot.

Solution: The idea there (and here) is that we want all the explanatory variable values in *one* column, so that we can plot them all on one plot. We did this in C32 by using `pivot_longer` to put them all in one column (while keeping track of which x-variable each value was):

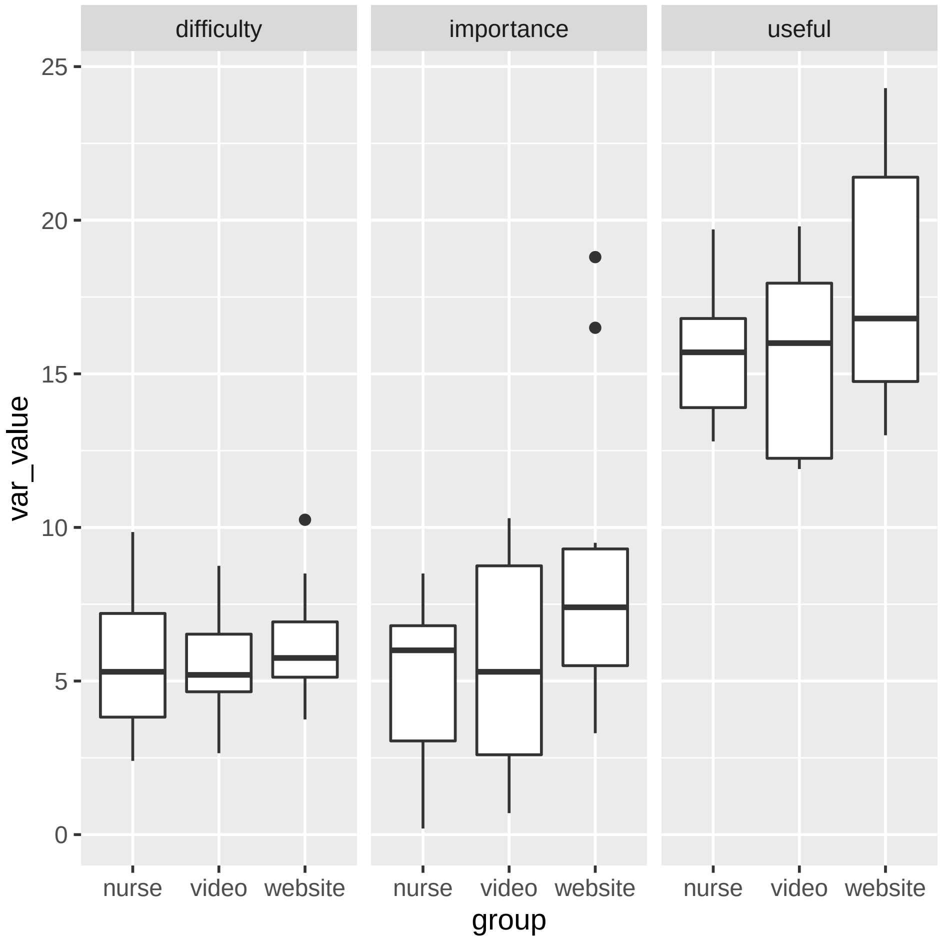
```
dietary %>% pivot_longer(-group, names_to="var_name",
                                 values_to="var_value") -> d
d
```

```
## # A tibble: 99 x 3
##    group   var_name   var_value
##    <chr>   <chr>          <dbl>
##  1 website useful         19.6
##  2 website difficulty      5.15
##  3 website importance      9.5
##  4 website useful         15.4
##  5 website difficulty      5.75
##  6 website importance      9.10
##  7 website useful         22.3
##  8 website difficulty      4.35
##  9 website importance      3.30
## 10 website useful         24.3
## # ... with 89 more rows
```

This makes a column `var_value` that has all the response variable values in it, and a column `var_name` that says which response variable it was.

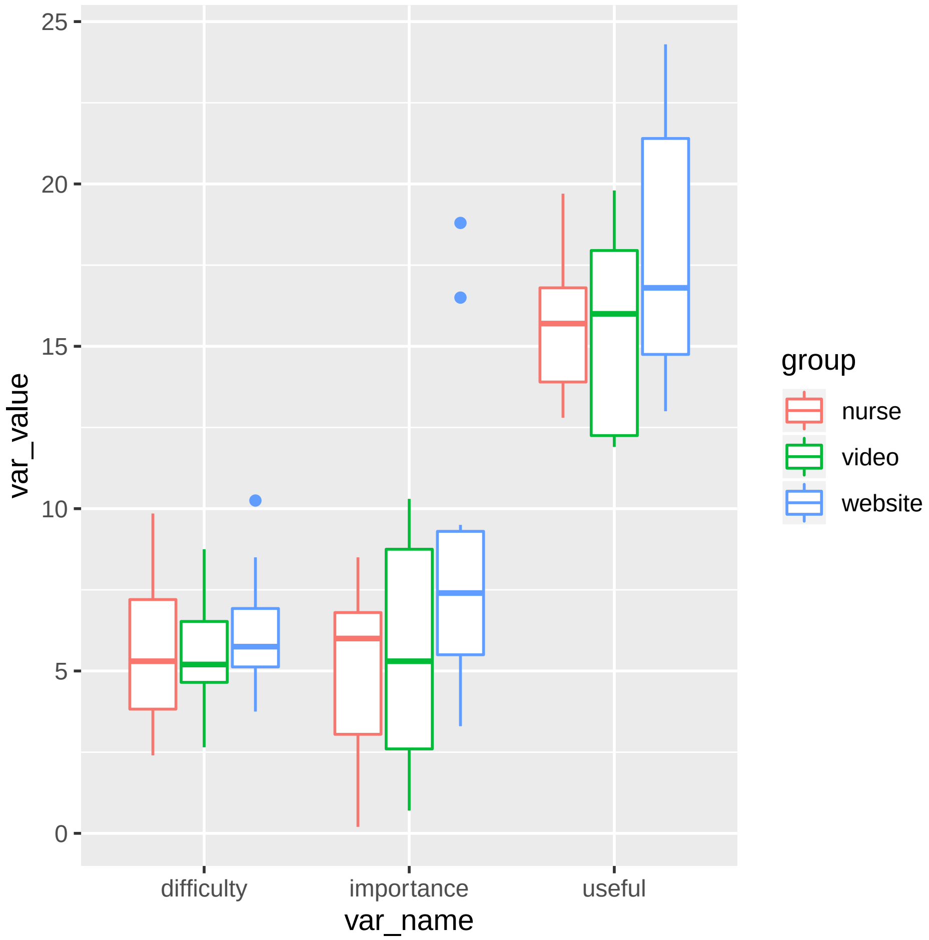
Now you have two ways you can go. The first way I thought of was to put the variable names in facets, thus:

```
ggplot(d, aes(x=group, y=var_value)) + geom_boxplot() +
    facet_wrap(~var_name)
```



but you could also take the view that variable name is a second categorical variable, and do grouped boxplots:

```
ggplot(d, aes(x=var_name, y=var_value, colour=group)) +
    geom_boxplot()
```

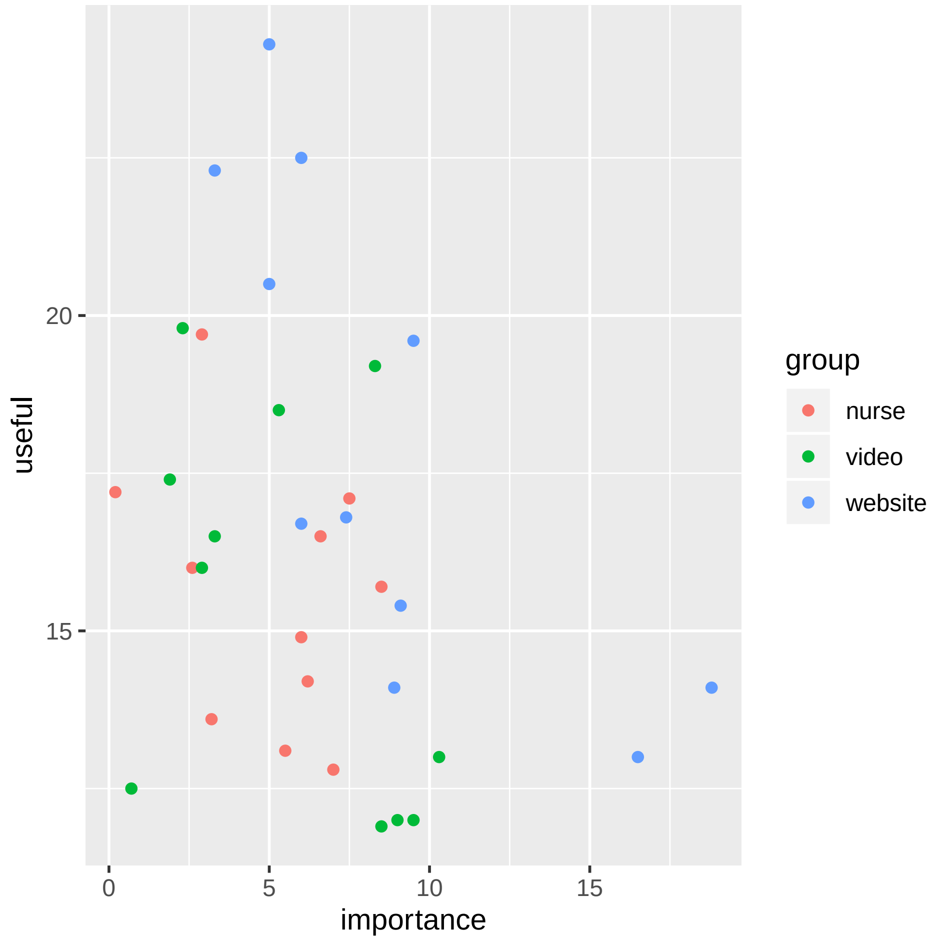


I had to think a bit, in the last one, about what I wanted to be `x` and what I wanted as `colour`. I decided in the end I wanted to compare group *within* variable, so I made the variable names be `x` and made the groups be `colour`.

If you couldn't manage the `pivot_longer`, make three boxplots, one after the other, for each response variable. This is not quite so good, because it doesn't put things side by side for comparison, so I call that 3 out of 4.

Extra: given that `difficulty` seems to be the same no matter which `group` a respondent was in, maybe we can think about the joint effect on `importance` and `useful` of `group`. This is something we can plot, because there are now only two quantitative variables, so a scatterplot with the `group` categories indicated by colour would do it:

```
ggplot(dietary, aes(x=importance, y=useful, colour=group)) + geom_point()
```



The blue dots are for the most part the ones most up and to the right, so it looks as if both `useful` and `importance` are typically highest for people who viewed the website. Or, perhaps more accurately, people who viewed the nurse's presentation, either live or on video, are at least sometimes low on one or both of `useful` and `importance`.

(d) (2 marks) What does your plot suggest to you about the research objectives mentioned in the opening paragraph? Explain briefly.

Solution: The first idea is to compare the three groups on each of the three response variables. I would say all three groups had about the same rating for `difficulty`, but on the other two variables, the group of people viewing the website had slightly higher ratings than the other groups. Maybe `importance` shows the biggest effect. This is a matter of opinion, so have one, even if it's not the same as mine. If the plot tells the story you tell, it's good.

(e) (4 marks) Run a suitable MANOVA on these data. Think about what your response variables are, what you have to do with them, what your explanatory variable(s) is/are, and how to run the analysis. There are several steps. What do you conclude in the end?

Solution: There are three response variables, `difficulty`, `importance` and `useful`. There is one explanatory variable `group` which is indeed categorical.

The steps are:

- Make a response matrix out of the response variables.
- Fit using `manova` or `lm` (depending how you're going to look at it next)
- Look at the results using `summary` (if you used `manova`) or `Manova` from package `car` (if you used `lm`).

Making the response is the same either way:

```
response <- with(dietary, cbind(difficulty, importance, useful))
```

Or, omit the `with` and use *three* dollar signs (to say that each of the three columns in turn is from data frame `dietary` or whatever you called it).

That looks like this (which you don't have to display, but I want to talk about it):

```
response
```

```
##       difficulty importance useful
##  [1,]       5.15        9.5   19.6
##  [2,]       5.75        9.1   15.4
##  [3,]       4.35        3.3   22.3
##  [4,]       7.55        5.0   24.3
##  [5,]       8.50        6.0   22.5
##  [6,]      10.25        5.0   20.5
##  [7,]       5.95       18.8   14.1
##  [8,]       6.30       16.5   13.0
##  [9,]       5.45        8.9   14.1
## [10,]       3.75        6.0   16.7
## [11,]       5.10        7.4   16.8
## [12,]       9.00        7.5   17.1
## [13,]       5.30        8.5   15.7
## [14,]       9.85        6.0   14.9
## [15,]       3.60        2.9   19.7
## [16,]       4.05        0.2   17.2
## [17,]       4.40        2.6   16.0
## [18,]       7.15        7.0   12.8
## [19,]       7.25        3.2   13.6
## [20,]       5.30        6.2   14.2
## [21,]       3.10        5.5   13.1
## [22,]       2.40        6.6   16.5
## [23,]       4.55        2.9   16.0
## [24,]       2.65        0.7   12.5
## [25,]       6.50        5.3   18.5
## [26,]       4.85        8.3   19.2
## [27,]       8.75        9.0   12.0
## [28,]       5.20       10.3   13.0
## [29,]       4.75        8.5   11.9
## [30,]       5.85        9.5   12.0
## [31,]       2.85        2.3   19.8
## [32,]       6.55        3.3   16.5
## [33,]       6.60        1.9   17.4
```

The square brackets on the left indicate that this is an R `matrix`, which is what we need here. `cbind`, when given vectors, will create a `matrix`; if you were to use `bind_cols` from the `tidyverse`, you'll get a `tibble` which will *not* work here. It's a matter of knowing what you want.

Then, choice of two ways. I think the first one is easier:

```
dietary.1 <- manova(response~group, data=dietary)
summary(dietary.1)
```

```
##           Df  Pillai approx F num Df den Df  Pr(>F)
## group      2 0.47667   3.0248      6     58 0.01215 *
## Residuals 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can also do it this way, as long as you have a `library(car)` somewhere:

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
```

```
dietary.2 <- lm(response~group, data=dietary)
Manova(dietary.2)
```

```
##
## Type II MANOVA Tests: Pillai test statistic
##       Df test stat approx F num Df den Df  Pr(>F)
## group  2   0.47667   3.0248      6     58 0.01215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Whether you use small-m `manova` or big-M `Manova` is entirely up to you. The answer will be the same either way. The value of the second way is that it's the way of doing repeated measures, so if you go that way here you have a bit less to learn when you're doing that.

Let's recall our null hypothesis for this kind of thing: the means of each response variable are the same for each group. That is, the mean of `difficulty` is the same for each `group`, the mean of `importance` is the same for each group (but not necessarily the same as the mean of `difficulty` was), and so on. We are going to reject this null hypothesis, but doing so leaves us in an even worse position than it does with ANOVA: at least one of the response variables differs in at least one of the groups.

Big hurrah, right? Unfortunately, this is as far as we can go. We can eyeball our boxplots for some insight as to why it came out this way, or we can wait (until after repeated measures) for discriminant analysis which is a way to understand MANOVA results.

Extra: if you go the `Manova` way, you can also do this:

```
summary(Anova(dietary.2))
```

```
##
## Type II MANOVA Tests:
##
## Sum of squares and products for error:
##            difficulty importance    useful
## difficulty 126.287277   34.18591  6.550907
## importance  34.185908  426.37090 -207.777259
## useful       6.550907 -207.77726  293.965442
##
## ------------------------------------------
##
## Term: group
##
## Sum of squares and products for the hypothesis:
##            difficulty importance   useful
## difficulty   3.975151   16.71121 14.24394
## importance  16.711210   81.82969 64.55151
## useful      14.243938   64.55151 52.92424
##
## Multivariate Tests: group
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            2 0.4766701 3.024828      6     58 0.01215223 *
## Wilks             2 0.5257884 3.538230      6     56 0.00485936 **
## Hotelling-Lawley  2 0.8972300 4.037535      6     54 0.00205762 **
## Roy               2 0.8919879 8.622550      3     29 0.00030233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and you can check the consistency of P-values from the four tests given there. The one we had before, which is called Pillai's Trace, actually has the largest P-value of all of them. You are looking for some kind of consensus; here all of them are clear rejections of the null, with the (non-)conclusion we had above.

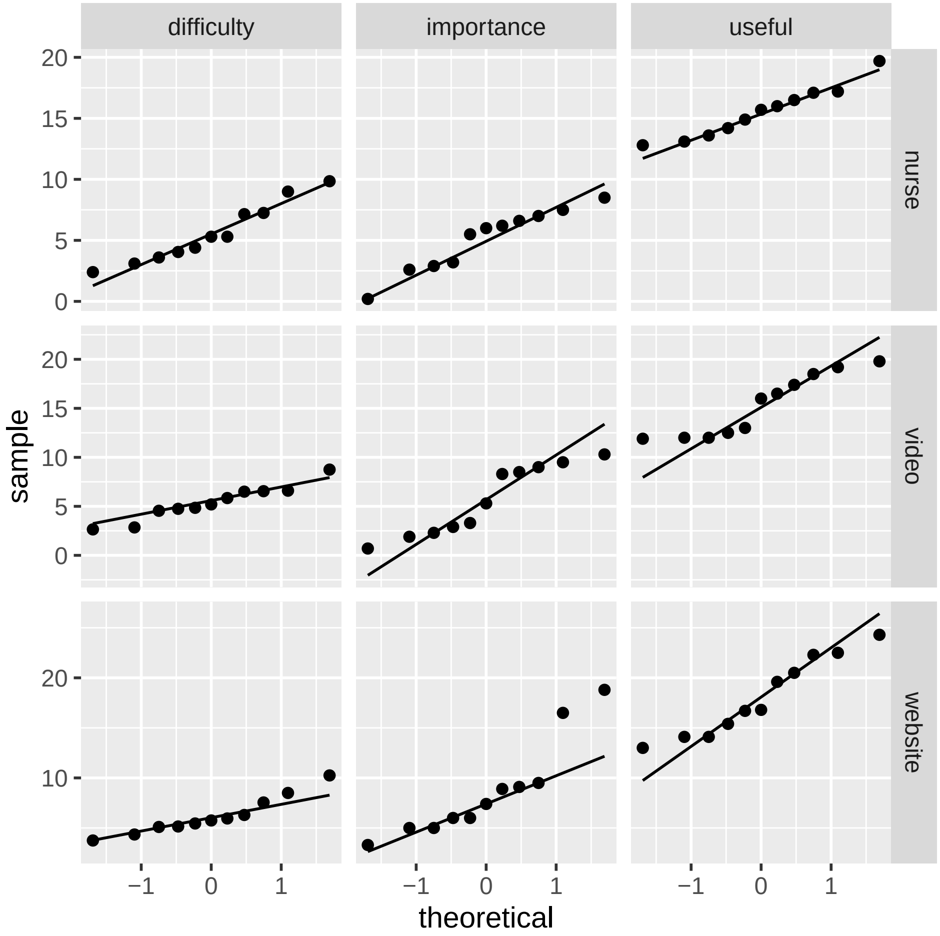
Extra 2: there are two assumptions lurking in the background here, analogous to the normality and equal-variances assumptions of regular ANOVA. These are:

- multivariate normality of data within each group

- equality of variances *and* correlations within each group

Multivariate normality is tricky to test for. One way that gets at least at part of it is to test the normality of each response variable within each group. This has a fancy facetted solution like our facetted boxplots earlier:

```
dietary %>% pivot_longer(-group, names_to="var_name",
                                 values_to="var_value") %>%
  ggplot(aes(sample=var_value)) +
    stat_qq() + stat_qq_line() +
    facet_grid(group~var_name, scales="free")
```



The problem here is that a normal quantile plot has only one variable (the quantitative one whose normality we are testing), so we now have two things to facet by: the variable name and the groups. `facet_grid` does a two-dimensional facetting; the thing before the squiggle is the y-scale for the facets, and the thing after the squiggle is the x-scale for the facets. Hence, as I did it, the groups go up and down, and the variable names go left and right. The idea is that all nine plots should be approximately normal. What do you think?

For me, all of these are OK except for the importance-website one, that has two upper outliers. (I'm not bothered by the other ones with *short* tails, since they won't affect the mean.) These showed up on the boxplot also.

The problem with rejecting (multivariate) normality in this context is that we don't know what to do instead. I'm thinking that there ought to be a multivariate version of Mood's Median Test, where you count the number of values above and below the overall median for each variable and for each group, so you get kind of a stack of tables, one above each other, and then you attack that by something called a log-linear model, which we see at the end of this course. But I don't know if that even works.

The second part is testing equality of covariance matrices. This can be done by a thing called Box's M test. I know you know my feelings towards doing tests to see whether you can do a test, and indeed George Box himself had similar misgivings:

> To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave

> port. — G. E. P. Box (1953)

but, anyway, if you're OK enough with multivariate normality, which I'm going to pretend we are here, Box's M test lives in a package called `heplots` which you'll have to install first, and then:

```
library(heplots)
boxM(response~group, data=dietary)
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 10.286, df = 12, p-value = 0.5909
```

This clearly fails to reject equal covariance matrices, and so if you're OK with the multivariate normality, the MANOVA is to be trusted. This test is actually *very* sensitive; the advice I've seen is to be OK with equal covariance matrices even if the P-value gets down to 0.001, but not below that.