University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 6, 2017, 7:00–10:00pm (3 hours)

Aids allowed:

- My lecture overheads (slides)
- The R “book”
- Any notes that you have taken in this course
- Your marked assignments
- Non-programmable, non-communicating calculator

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 26 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker’s attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and in the table on the next page.

The University of Toronto’s Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Last name: ____

First name: ____

Student number: ____

For marker's use only:

Page	Points	Score
1	6	
2	4	
3	4	
4	10	
6	7	
7	4	
8	8	
10	5	
11	6	
14	6	
15	2	
16	11	
17	2	
19	3	
22	6	
24	8	
25	9	
Total:	101	

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

1. American football is a tough physical game, and players wear helmets to protect their heads. There is concern, however, about a possible association between helmet design and neck injuries. As a preliminary to a study of this association, data were collected on three groups of individuals, labelled `group` in the data: high school football players (labelled 1 in the data set), college football players (labelled 2) and non-football players (3). For each individual, six head measurements were made, as follows:

   - `wdim`: head width (at its widest)
   - `circum`: head circumference
   - `fbeye`: front-to-back head size at eye level
   - `eyehd`: eye to top of head measurement
   - `earhd`: ear to top of head measurement
   - `jaw`: jaw width.

   All of the measurements are in inches.

   Some randomly-chosen rows of the data are shown in Figure 2. Our aim is to predict which one of the three groups an individual is in, based on their head measurements.

   (a) (2 marks) What two things about the data suggest that I should use `multinom` for the modelling?

   My answer: The response variable `group` is (i) a factor (categorical variable), and (ii) its levels are unordered (1, 2, and 3 serve only to identify the three groups, but they have no meaning as numbers).

   Try to keep distinct the number of *variables* (`group`, a factor, plus the six explanatory variables), and the number of *levels* of a factor (three, the three groups).

   Also, if you use the word "nominal", you should explain what it means, or at least give some kind of hint that you know what it means.

   (b) (2 marks) Some modelling is shown in Figure 3. Explain briefly in words what the `update` line is doing.

   My answer: Take the model `football.1` (which has all the explanatory variables in it) and, from it, remove the two variables `circum` and `fbeye`, and call that model `football.2`.

   I want to see a word like "remove", since that is what this `update` statement is doing. The `anova` below shows you which two models are being compared, so that you can see that the second model has four explanatory variables in it, and therefore `football.2` must be that model, but the important point here is *how* model `football.2` got to contain the explanatory variables it did, rather than which ones it contains (which you can infer from the `anova`).

   (c) (2 marks) What do you conclude, in the context of the data, from the test done in Figure 3? Explain briefly.

   My answer: By this, I mean the `anova` at the bottom of the output. This is comparing the two models, one with all the explanatory variables, and the other without the two variables that were removed. The P-value is not small, so there is no significant difference (in fit) between the two models, and therefore we should go with the smaller model `football.2` because it is simpler,

(d) (3 marks) I did some predictions for the model `football.2`. (You may assume that this is a reasonable model to be predicting from.) These are shown in Figure 4. Using the information in this Figure, what can you say about an individual who has head width (at its widest) 15, eye-to-top-of-head 10, ear-to-top-of-head 13.5, and jaw width 12.5 inches?

> **My answer:** Find this individual in the predictions (the variables are in the right order). This is row 9. The last three columns are the predicted probabilities of an individual with these measurements being in each of the three groups: probability 0.12 of being a high school football player, probability 0.72 of being a college football player, and probability 0.16 of not being a football player at all.
>
> For full credit, you need to identify what the 1, 2, and 3 *mean*. It is not enough to say “the probability of being in group 1”. You probably need to identify at least one of (i) the row of the predictions (row 9) or (ii) the actual probabilities. If you just say “this is most likely a college football player”, I don’t know that you’re looking at the right row, since there are other rows where the probability of being in group 2 is highest.
>
> In the spirit of discriminant analysis, the highest predicted probability is the one going with “college football player”, so, if you had to guess which group this individual belonged in, you would guess they were a college football player.
>
> There are no actual injuries here (that study came later). All we’re doing here is trying to predict which group the player is in, based on his measurements.

(e) (1 mark) The data for this problem had one individual per row of the data file. Suppose, instead, each row represented several individuals, with a column called `count` saying how many individuals were summarized in that row. How would I have to change my `multinom` statement to accommodate this?

> **My answer:** Add `weights=count` to it.
>
> I know, this is almost asking you to write code, which I said I would not ask you to do, but I think it is really a modelling issue rather than a coding one.
>
> If you manage to come up with something else that will work, I’ll go for that (though I don’t think anyone did). The best possibility I can think of is to make the right number of copies of each row, according to the value in `count`.
>
> I was willing to go for something like `weights=freq`, on the basis that if you did that in front of a computer, you’d soon discover what was wrong, but you need to have exactly `weights` in that slot.

2. The “windshield” method of distributing questionnaires consists of randomly placing the questionnaires under the windshield wipers of cars in parking lots. What method leads to the best response rate (that is, the fraction of questionnaires filled out and returned)? One idea is to use brightly-coloured paper to print the questionnaires on. In one study, blue, green and orange paper was used. 15 supermarket parking lots in a metropolitan area were chosen. One colour was randomly chosen for each parking lot, and questionnaires of that colour were distributed. Does the colour of the paper make a difference to the response rate? One of the researchers thought the size of the parking lot (number of parking spaces contained in the entire lot) might also matter, so this was recorded as well. The data are shown in Figure 5. The response rate is in percent.

   A scatterplot of the data is shown in Figure 35 (at the end of the booklet of code and output, this plot being in colour), with response rate plotted against parking lot size, with the points coloured according to the colour of the questionnaire paper and regression lines added for each colour. (The purpose of the last line of code above the plot is to make the colours on the plot correspond to the colours of the paper. Otherwise the line labelled “blue” would actually have been coloured red!)

(a) (2 marks) Looking at Figure 35, how does it appear that parking lot size affects the response rate, if at all? Explain briefly.

My answer: The lines all go downhill, so that for a bigger lot, the response rate is lower (and this appears to be true no matter what the colour of the questionnaire).

(b) (2 marks) Again looking at Figure 35, does it appear that the colour of the paper the questionnaire is printed on has any effect on the response rate, or not? Explain briefly.

My answer: The green points are consistently highest (and the orange ones consistently lowest), for any fixed size of parking lot, so that the response rate is highest for green paper and lowest for orange paper: that is, the colour of the paper does appear to make a difference to the response rate. Or you can pick a parking lot size, say 300, note that green has the highest response rate and orange the lowest, and then observe that this same colour pattern holds no matter what parking lot size you pick.

This one is very easy to overthink. It doesn't have anything to do with interaction (coming up later).

(c) (2 marks) In Figure 6, two models are fit and compared. Explain briefly why I can remove the interaction between size and colour, based on the information in this Figure.

> **My answer:** The models being compared have both main effects and the interaction (model `qq.1`) and just the main effects without the interaction (`qq.2`). The `anova` says there is no significant difference between these two models, so we prefer the simpler one, that is, `qq.2` without the interaction.
>
> The best way to tackle this one is the usual `anova` way. The test isn't in itself a test of the interaction (it just happens to work out this way because of the models being compared).

(d) (2 marks) What does the model without interaction mean for these data?

> **My answer:** The key is "the lines are parallel". Specifically, the lines for the relationship between response rate and size are parallel (have the same slope) for each colour paper.
>
> If you didn't see that this was analysis of covariance (though the picture with the almost-parallel lines was meant to be a clue), you can also interpret a non-interaction the usual ways: the lot size and colour act independently on response rate, the response rate does not depend on the combination of colour and lot size, the effect of lot size is the same for each colour (or the effect of colour is the same for all lot sizes), any of those. I wanted to see at least something of this kind, though.

(e) (3 marks) Interpret each of the `size`, `colourgreen` and `colourorange` estimates in Figure 7. That is, explain what each of those three numbers mean, in the context of the data.

> **My answer:** The `size` estimate, -0.0298, says that as parking lot size goes up by 1, estimated response rate goes down by 0.03. Or, to put it on a better scale, if parking lot size goes up by 100, response rate goes down by 3 percentage points. This is a genuine slope; it's actually the slope of all three lines, since the lack of interaction implies that all three lines are parallel, ie. have the same slope.
>
> The two `colour` estimates compare the colours named to the baseline colour blue. The `colourgreen` estimate of 1.35 says that questionnaires printed on green paper have a response rate 1.35 percentage points higher on average than those printed on blue, all else being equal (that is, when the questionnaires are distributed in parking lots of the same size).
>
> `colourorange` is -1.77, saying that the response rate for questionnaires printed on orange paper is on average 1.77 percentage points *lower* than for those printed on blue, again, with parking lot size being equal, or "adjusting for parking lot size" if you prefer.
>
> I wanted you to tell me what those actual numbers meant (that is, to go further than saying whether they were positive or negative, or that they were significantly nonzero, which wasn't the point here). When you're looking at `colourgreen` or `colourorange`, you need to say what the increase or decrease is compared with (the baseline colour blue). So you need to pull out the three *actual estimates* and tell me exactly what the numerical values mean. Not mentioning the actual numbers will cost you a point, even if everything else is correct. Not mentioning that you are comparing with the baseline blue will cost you a point. Talking about something true but irrelevant (that shows *some* insight) *might* get you 1 out of 3.
>
> This is a rare care where it was possible to *underthink* your answer. I wanted to know what the actual *numbers* meant, and I couldn't think of a clearer way to say that.

(f) (3 marks) Figure 8 shows another analysis. What does this one show, and how, despite appearances, is it consistent with what you concluded earlier? Explain briefly.

My answer: This says that there is no difference in mean response rates between questionnaires of the different colours. (1 point for this.)

This appears to be in direct contradiction to our analysis of covariance, where colour of questionnaire *did* make a difference. However, what the analysis of covariance actually said was that *once you allow for the size of the parking lot*, colour of questionnaire makes a difference. The ANOVA in Figure 8 does not contain `size`, so it is not allowing for size. (Or, the `size` term in the ANCOVA was strongly significant, which means that taking it out, as the ANOVA implicitly does, was a mistake.) We should therefore trust Figure 7 and *not* Figure 8.

Or, Figure 7 *adjusts* for parking lot size, and Figure 8 does not. We need to adjust, because parking lot size makes a (big) difference. Failing to adjust for `size` is a mistake.

Two points for saying one of these, or something along the lines of it being a mistake to not include `size`. Mentioning `size` in some fashion was likely to be worth something.

There are different ways to say this. If your way seems to get at the issues, I'm happy with it.

3. Two treatments, labelled A and B, for post-traumatic stress disorder (PTSD) are being compared. Nine patients are available; each one is randomly allocated to one of the two treatments, labelled `trt` in the data. The two treatments include different amounts of supportive counselling and other therapies. The response variable is the number of symptoms of PTSD observed by the therapist (so a smaller number is better).

   The research questions are:

   1. Are the treatments doing something (changing the number of symptoms over time, ideally reducing the number)?
   2. Is one treatment more effective than the other?
   3. Is the pattern of change in symptoms over time different for the two treatments?

   Each patient was observed on three occasions: before therapy (`pre`), after therapy (`post`), and three months after therapy ended (`followup`). The data are shown in Figure 9.

   (a) (2 marks) Why is a repeated-measures analysis necessary here? Explain briefly.

   My answer: Each patient was observed not once but at multiple (3) times, so that we have three measurements on the same individuals. The three measurements on the same patient are likely to be correlated, because some patients will tend to have a lot of symptoms and some not so many.

   Strictly speaking, we need the *same* thing (number of symptoms, here) being measured at *different* times. If it was different things being measured, we'd have something like a regular MANOVA instead. But I thought it was mean to take off a point for saying something like "three different responses", so I didn't.

   Contrasting with ANOVA (one measurement per subject) was a nice way to show me that you knew what you were talking about.

   (b) (3 marks) A repeated-measures analysis of variance is shown in Figure 10. From this analysis, what do you conclude about each of the three research questions given above, in the context of the data?

   My answer: I should probably start with the interaction: this is not significant, which means that the pattern of symptom count over time is the same for both treatments. This is research question 3.

   Having found a non-significant interaction, in this kind of analysis we go ahead and look at the main effects: there is a significant effect of time, meaning that the symptom counts are changing over time. Whether they are going up or down we can't say yet. This is research question 1.

   Finally, there is no difference between the treatments (the test for `trt`), so the treatments are equally effective on average (research question 2).

   I have no objection to the order in which you take these; in particular, there's no obligation here to start with the interaction test. Here, it seemed logical to list the research questions in the order I did, and it seems equally logical to answer them in the same order. I would, however, like to see that you know which test answers which research question. There was a little trouble with research question 1, which is actually asking for a *time* effect, and the second question is answered by the treatment effect (not signficant), which says that whatever the treatments are doing over time is the *same* for the two treatments. (Not that the treatments are both doing nothing: that would be the *time* effect, and possibly the interaction; see discussion below.)

   (c) (2 marks) A spaghetti plot is shown in Figure 36 (at the end of the booklet of code and output). In the code above the plot, why did I need the `mutate` line? (Think about what would have happened if I had left it out.)

My answer: If I had used the original `time` column, the three levels of the factor would have appeared on the graph in *alphabetical order*, which would have put `followup` first and `pre` last, exactly backwards from the logical order! So I created an ordered factor, with the time points in the order I wanted them, and used that in the plot instead of `time`. That way, the time points came out in the right order.

I didn't need you to observe that the time points would have come out in alphabetical order, though it was nice if you did (this part was to guide your thinking). "To make the time points come out in the right order" was enough.

(d) (4 marks) How does the spaghetti plot support each of your conclusions of part (b)? In addition, does the spaghetti plot suggest that the treatments have a beneficial effect? ("Beneficial" means "does good".) Explain briefly.

My answer: The lack of interaction shows in the patterns of symptoms over time are about the same for the patients on each of the treatments: the red and blue traces go down in about the same way.

The lack of treatment effect shows in the red and blue traces being all intermingled: the red ones are not systematically better than the blue ones, for example.

The time effect shows in the traces generally going down as we go from before the treatment to after it.

This last point, that the traces are going *down* over time rather than up, is how we can tell that the treatments are both effective, since they both reduce the number of symptoms over time, at least on average. I didn't insist on your explicitly saying this, as long as you said somewhere that the number of symptoms typically went down over time (and if you observed that it went back up after `post`, it didn't go up as much).

This was a relatively simple example of this type of problem. One of the ways in which one of the treatments could have been more effective than the other is to reduce the number of symptoms to the same level in the end, but to do it *more quickly*. On the spaghetti plot, that would have shown up in the numbers of symptoms for both treatments being similar at `pre` and `followup`, but the more effective treatment would have fewer symptoms at `post`. Because the *pattern* of symptoms is different over time for the two treatments, this would show up in a significant *treatment by time interaction*. What you would then do is to draw the spaghetti plot and try to understand what that significant interaction meant: in the case I described, that one of the treatments is actually more effective than the other. We've seen this kind of thing before, so I didn't want you to have to grapple with it on the final exam, and therefore I gave you a simpler one.

4. In each of the cases below, describe specifically how the *data* appropriate for the two techniques given would be different.

   (a) (2 marks) Logistic regression using code as in Figure 11, and using `polr` from package `MASS`.

My answer: The response variable is categorical in both cases; in the first case, it has exactly two categories (levels), in the second case it has more than two, and the categories come in a natural order.

The difference is "the number of levels of the response", but also, in the case of `polr`, the nature of those levels (ordered vs. unordered). When you have only two levels, it doesn't matter whether they are ordered or not, because one is first and the other is second, either way.

Referring to "the data" rather than "the response variable" is not precise enough, because the variables on the *right* side of the squiggle can be anything.

Also, I prefer you to use some term other than "ordinal response", since it's too easy to pull that out of your notes and give it as an answer without knowing what it means. So say "a categorical variable whose levels come in a natural order", and then it's clear (i) what you mean and (ii) that you understand what you are saying. (I was willing to let this go if there was some other evidence that you knew what you meant.)

Here, again, you can get tripped by not being clear about the distinction between "several response variables" and "one response variable with multiple levels".

   (b) (2 marks) ANOVA and MANOVA. (You need to know what these two abbreviations mean.)

My answer: ANOVA is regular analysis of variance, as in C32, with one response variable (and any number of categorical explanatory variables), while MANOVA, multivariate analysis of variance, has more than one response variable, varying together.

Thus the difference is the number of response variables, ANOVA having one and MANOVA more than one.

This is all you needed, but a number of people managed to make it a lot more complicated than this. Repeated measures (as we have done it, ie. not the `lmer` stuff) is a kind of MANOVA, but only one kind; there are other MANOVAs, with multiple genuinely different response variables (like the Summit Creek data coming up later), the kind of thing for which discriminant analysis is a followup. Also, don't get confused between the R function `anova` that is used for comparing two models (or obtaining an analysis of variance for a single one); I meant ANOVA the statistical technique, for which we used `aov` followed by Tukey.

   (c) (2 marks) Analysis of variance and analysis of covariance.

My answer: This time, the difference is in the explanatory variables. Analysis of variance has only categorical explanatory variables (one or two, as we have seen it), while analysis of covariance has one (or more) quantitative explanatory variables in addition. The typical ANCOVA we have seen has one quantitative and one categorical explanatory variable, but it could have more.

Thus the difference is categorical-only explanatory variables vs. one (or more) numerical ones *in addition.*

Most people figured this one out. One way to get yourself confused here is to think only of ANOVA as looking for differences between groups, which it is, but it doesn't help in thinking about what kind of variables you have, which is the key to this one. Some people thought that interactions were the difference, but you can have interactions in both 2-way ANOVA and in ANCOVA, with the additional interpretation in ANCOVA of lines being parallel or not.

   (d) (2 marks) Hierarchical clustering, as with `hclust`, and K-means clustering.

My answer: As we have seen it, hierarchical clustering needs distances or dissimilarities, while K-means uses actual data on variables.

It's not really as clear-cut as that, since you can always start with data on variables and make distances from them (and then use `hclust`) if you want to do that, but in terms of the input data the methods use, dissimilarities vs. actual data is the distinction.

I thought this was the best answer, since it gets at how the *data* would be different for the two methods, but some credit is available for other relevant observations, such as: `hclust` lets you figure out how many clusters and `kmeans` makes you say up front how many you want, or: `hclust` gives you the whole clustering process and you can "chop" it where you want, while `kmeans` only gives you the number of clusters you asked for.

This was much the most challenging of the four parts.

5. In this question, I will show you some examples of R code and ask you to explain *in words* what each code example does.

   (a) (3 marks) Figure 12 shows a data frame `x` and some code, on the last line. The results of the pipe on the last line are saved in a variable `y`. What will `y` look like?

My answer: Here's the code again, this time displaying the result:

```
x=data.frame(id=1:2,t1=c(10,11),t2=c(12,14),t3=c(13,16))
x
##   id t1 t2 t3
## 1  1 10 12 13
## 2  2 11 14 16
x %>% gather(time,resp,t1:t3) -> y
y
##   id time resp
## 1  1   t1   10
## 2  2   t1   11
## 3  1   t2   12
## 4  2   t2   14
## 5  1   t3   13
## 6  2   t3   16
```

This is the standard way to turn "wide format" (`x`) into "long format" (`y`). There will be a column `resp` with all six responses (values for two different people at three different times), and a column indicating which time the response value came from. There is also a column `id` that is repeated three times, once for each time, so that each response value is identified by which `id` and which time it belongs to.

It's easy for me to check that you have the right thing by listing (some of) `y`, but you don't have to. Conversely, you can give just the answer, but you help yourself by adding some explanation (in case something goes wrong in your answer; I can give you partial credit if there is something useful in your explanation).

Using the word "gather" in your explanation is not very helpful, because you are trying to *explain* what `gather` does, and explaining `gather` with "gather" looks as if you are guessing.

This was a very standard application of `gather` to turn wide format into long format, so I would expect you to recognize it by now. There were quite a few very good answers.

   (b) (2 marks) Using the same data frame `x` as above, look at Figure 13. What does `w` contain?

My answer: Once again, showing the answer as well:

```
x
##   id t1 t2 t3
## 1  1 10 12 13
## 2  2 11 14 16
w=map_dbl(x,mean)
w
##   id   t1   t2   t3
##  1.5 10.5 13.0 14.5
```

This calculates the mean *for each column of x*. Here, of course, it makes no sense to calculate the mean `id`, so that in practice you would de-`select` that column first, to get the means of only the other columns. Thus I accepted an answer "mean response for each time" or "mean of columns `t1` through `t3`", since that is what it would make sense to obtain, but actually getting that would have been a little messier:

```
x %>% select(-id) %>% map_dbl(mean)
##   t1   t2   t3
## 10.5 13.0 14.5
```

The way we know how to get means of things is to make the "things" rows, and then use `group_by` and `summarize`, so this would do the same thing, just formatted differently:

```
x %>% gather(time,resp,t1:t3) %>%
  group_by(time) %>%
  summarize(tmean=mean(resp))
## # A tibble: 3 x 2
##   time  tmean
##   <chr> <dbl>
## 1 t1     10.5
## 2 t2     13
## 3 t3     14.5
```

The programming language Perl has an acronym TMTOWTDI, which stands for "there's more than one way to do it". R is like that too.

(c) (2 marks) Look at Figure 14. Starting from the vector `xx` as shown, what does the vector `yy` contain? (We have used the function `sqrt` in this course, so you need to know what it does.)

My answer: Here is the code again, this time displaying the answer:

```
xx=1:4
xx
## [1] 1 2 3 4
yy=map_dbl(xx,sqrt)
yy
## [1] 1.000000 1.414214 1.732051 2.000000
```

These are the square roots of the numbers 1 through 4. The logic of the `map` functions is "for each of the first thing, do the second thing".

I was fairly generous about what I would take, but I drew the line at "the square root of `xx`", unless you said what that meant, since `xx` is a vector. You needed to say or at least imply that you wanted the square root of *each* value.

R knows what the square root of a vector is, actually:

```
sqrt(xx)
## [1] 1.000000 1.414214 1.732051 2.000000
```

but `map_dbl` also applies to a "non-vectorized" function, that is, one that *only* produces a number from its imput, even if its input is a vector. (This was, therefore, an artificially simple example, but I didn't want to confuse things unnecessarily.)

(d) (4 marks) Look at Figure 15. There, a function `f` is defined, a data frame `data` is shown, and a last line of code saves something into a variable `res`. What, in words, does the function `f` do, and therefore what does `res` contain? For full credit, explain how `res` is laid out.

My answer: Here's the code again, plus some extra showing you how I generated data frame `data`:

```
set.seed(457299)
data=data.frame(x1=rnorm(10),x2=rnorm(10),x3=rnorm(10))
data
##                x1           x2           x3
## 1   1.621867352 -1.65547514  1.16077867
## 2  -0.746347365 -1.20687430  0.47187652
## 3  -0.268930797  1.26874912  0.94460805
## 4  -0.699535090  0.83839323 -0.80725768
## 5   0.213237930 -0.74610634  0.27918883
## 6   0.708968535  0.05275361  0.68644436
## 7  -1.078329045  1.51487539  0.60764160
## 8   0.791310415 -0.11230871  0.07134409
## 9   0.004046959  0.26653521 -0.15448600
## 10  1.095879569 -1.72037830 -1.17761202
f=function(mydata) {
  q1=quantile(mydata,0.25)
  q3=quantile(mydata,0.75)
  return(c(q1,q3))
}
f(1:9)
## 25% 75%
##   3   7
data
##                x1          x2          x3
## 1   1.621867352 -1.65547514  1.16077867
## 2  -0.746347365 -1.20687430  0.47187652
## 3  -0.268930797  1.26874912  0.94460805
## 4  -0.699535090  0.83839323 -0.80725768
## 5   0.213237930 -0.74610634  0.27918883
## 6   0.708968535  0.05275361  0.68644436
## 7  -1.078329045  1.51487539  0.60764160
## 8   0.791310415 -0.11230871  0.07134409
## 9   0.004046959  0.26653521 -0.15448600
## 10  1.095879569 -1.72037830 -1.17761202
data %>% map(f) %>% bind_rows() -> res
res
## # A tibble: 2 x 3
##        x1     x2      x3
##     <dbl>  <dbl>   <dbl>
## 1 -0.592 -1.09  -0.0980
## 2  0.771  0.695  0.667
```

The function calculates the first and third quartiles of a vector of data that you feed it (and returns those). One point. The last line calculates the first and third quartiles of *each column* of `data` and glues them back together into a data frame (`map` gives back a `list`), two points, with the first quartile in the first row and the third quartile in the second row. This last part is what will get you the fourth point.

If I hadn't done the `bind_rows`, this is how it would have looked:

```
data %>% map(f)
## $x1
##        25%        75%
## -0.5918840  0.7707249
##
## $x2
##        25%        75%
## -1.0916823  0.6954287
##
## $x3
##         25%         75%
## -0.09802847  0.66674367
```

The `bind_rows` makes a column for each variable and constructs a data frame out of them.

You might have been expecting `bind_rows` to glue the three things in the list together as *rows*, with the variable names as rows and Q1, Q3 as columns, but it doesn't work that way: what was columns before stays as columns. This is true whether you use `bind_rows` or `bind_cols`:

```
data %>% map(f) %>% bind_cols()
```

```
## # A tibble: 2 x 3
##       x1     x2      x3
##    <dbl>  <dbl>   <dbl>
## 1 -0.592 -1.09  -0.0980
## 2  0.771  0.695  0.667
```

The separate list items are always turned into *columns* when they are "bound" together. I would like the "names" on the individual list items to become row names on the final data frame (tibble), but that's not how it works (Hadley doesn't like row names). Indeed, the individual items in the list might not even *have* names, and if they became columns, what names would they have then? The original variables are guaranteed to have names, so they retain those names in the final data frame. I realize that you might get this wrong even though you have the logic correct, but sorry, that's the way it is.

There was a logic to this whole question (apart from the `gather` piece, which was meant to be a freebie). (b) was about applying a function to each column of a data frame, (c) was about applying a function to each element of a vector, and (d) combined those ideas to apply a user-defined function, one that returns two values, to each column of a data frame, and then combined all the values into a usable thing. Thus, I was hoping that you would be able to guess that (d) would be finding the quartiles for each column.

If you remember back to C32, you might recall that the quartiles of the standard normal distribution, from which these data are randomly drawn, are these:

```
qnorm(c(0.25,0.75))
```

```
## [1] -0.6744898  0.6744898
```

The quartiles for our data are not very close to this, because we only have small samples.

6. The data set we use for this question consists of 88 observations of a number of variables taken along a 0.8km stretch of Summit Creek in eastern Oregon. The members of the research group that collected these data were interested in the effect of cattle grazing on the nature of the creek. They divided the creek into three zones, labelled `Reach` in the data set: an upstream area A in which cattle were permitted to graze, a middle area B from which cattle were excluded, and a downstream area C in which cattle were again permitted to graze. In each zone, the creek was divided into "hydrologic units" such as pools, shallow sections or straight sections, such that the variables measured (the widths and depths described below) would be expected to be similar within a hydrologic unit but possibly different for different hydrologic units.

   The variables measured for each hydrologic unit were these:

   - `DepthWS`: depth from water surface to bottom (metres)
   - `WidthWS`: width at water surface (metres)
   - `WidthBF`: width at "bankfull stage" (metres). This is a way of measuring the width of a creek that does not depend on how much water it contains. (That's all I know. I have no more information.)
   - `HUAreaWS`: area of hydrologic unit at water surface (square metres)
   - `HUAreaBF`: area of hydrologic unit at bankfull stage (square metres)
   - `wsgrad`: water-surface gradient (dimensionless).

   Some of the data is shown in Figure 16. There are other variables that will not be used in this question.

   (a) (2 marks) What do you conclude from the analysis in Figure 17 about the zones of the creek?

   > **My answer:** The zones (reaches) do not have the same means on all the variables; that is, one or more of the zones differ on one or more of the variables.
   >
   > More precise than that we cannot be, yet, which is why the discriminant analysis is coming.
   >
   > I want you to say something about zone not necessarily affecting *all* the response variables; it could be only some of them, or some combination of them (that we will discover in the discriminant analysis).
   >
   > In fact, the *zones* differ according to whether or not cattle grazing happens there, so we are not talking about cattle grazing (yet).

   (b) (2 marks) What would a discriminant analysis tell us about the creek that would be worth knowing? Explain briefly.

   > **My answer:** It will give us an idea of *how* the zones differ, now that we know that they do differ: that is, we will be able to see which if the variables help to distinguish the zones and how.
   >
   > This use of discriminant analysis is analogous to the use of Tukey after ANOVA: given that there is a difference between some groups, what kind of difference is it?

   (c) (2 marks) The variables are measured on very different scales, so I decided to standardize them before proceeding. The calculation is shown in Figure 18. The variables in the data frame `summit.s` have the same names as variables in data frame `summit`, but all the numeric variables in `summit.s` have been standardized. The standardized variables are used in the rest of the question. (This is for your information; there is no question here. The question for this part is in the short paragraph below.)

   Figure 19 shows the results of a discriminant analysis on the (standardized) data. Why are there two linear discriminants?

My answer: After all that preamble, this one is easy: 6 variables, 3 groups (zones), $\min(6, 3-1) = 2$.

You need to mention the number of variables somewhere (and get it right) for the second point; it's not just one less than the number of groups. (Think about what would happen if you just had one measured response variable.)

(d) (2 marks) Do you think both linear discriminants are worth paying attention to, or not? Explain briefly.

My answer: Look at the "proportion of trace" in Figure 19. The two values are 0.63 and 0.37. The second one is smaller, as it will be, but it is not *very* small compared to the first one, which suggests that both linear discriminants will have some value to us.

If you want to insist that the second proportion of trace value is "very small" compared to the first one, go ahead (and then you need to say that we should ignore `LD2`), but you might want to revisit that later when you look at the plots.

In the light of what I said above, the "proportion of trace" for `LD1` is *always* larger than for `LD2`, so this is not by itself a reason to ignore `LD2`. To justify ignoring `LD2`, you need to say that the proportion of trace is "much higher" for `LD1` or is "very small" for `LD2`, or something similar.

(e) (3 marks) Which of the variables in `summit.s` does the first linear discriminant mainly depend on? What would make a hydrological unit have a *large* score on `LD1`? Explain briefly.

> **My answer:** I would say that only `WidthWS` and `WidthBF` have any noticeable contribution to `LD1`, since the other coefficients are all near 0. I think this is clear, so I think this is the call I'd expect you to make. Having said that, if you said that the important variables were `WidthWS` only, or the first four variables (and not the last two), I couldn't very well take points off you, provided you followed through in your determination of when `LD1` would be large.
>
> Both of my two coefficients are positive, so a hydrological unit will have a large score on `LD1` if these two width measurements are *large*. If you included or omitted any variables in the first part of your answer, you ought to include or omit them here too. You ought to consider only the variables that you said were important, though I was fairly relaxed about this, as long as you said something not obviously nonsensical.

(f) (3 marks) Which three variables make the largest contribution to `LD2`, and what, therefore, would make a hydrological unit's `LD2` score *small*? Explain briefly.

> **My answer:** `WidthBF`, `HUAreaWS`, and `HUAreaBF` have the three biggest coefficients in size. The first and last of these are negative and the second one is positive, so the `LD2` score will be small if `WidthBF` is large, `HUAreaWS` is small, and `HUAreaBF` is small.
>
> Clues to the last two parts are on the biplot, Figure 37. The two width variables are about the only thing that point left and right, so they are the two things that belong to `LD1`, while the three variables mentioned in the previous paragraph are the ones that have the biggest influence up and down. `DepthWS` points more down than across, but it has a short arrow overall, so its contribution to `LD2` is smaller than that of `WidthBF`. If you want to use the biplot to support or confirm your answers above, go ahead.

(g) (3 marks) Zone B is the only one in which cattle are not allowed to graze. Looking at either or both of Figure 37 and Figure 38 (in colour at the end of the booklet of code and output), how is zone B different from the other zones? What does that imply about the values on the measured variables for hydrological units in zone B?

> **My answer:** Zone B, the green dots, is on the left side of the two Figures. That means the hydrological units there have low scores on `LD1`. Looking at either the biplot or the Coefficients of Linear Discriminants (or your earlier answer), zone B is low on `WidthWS` and `WidthBF`. Or you can get this from the Group Means in Figure 19.
>
> If you want to, you can also say that Zone B is typically also low on LD2, and therefore the variable values that go with a low LD2 score is what you will have. I think low LD2 is less consistent with the green dots than low LD1, but if you can make a consistent story I'm happy with it.
>
> Some people said that none of the arrows in the biplot are pointing at Zone B. This is a fair observation, but you have to be careful interpreting it: what it means is that Zone B is *low* on the variables whose arrows it is at the *tail* of (that is, that point directly *away* from zone B): particularly `WidthWS` and to a lesser extent `WidthBF`). To a small extent `DepthWS` points towards the Zone B points, and so a higher value on that might be associated with being in Zone B.

(h) (2 marks) Look again at Figure 37 and Figure 38. Is there an indication that `LD2` does anything to distinguish the zones? Explain briefly.

My answer: Most of the red dots (zone A) are at the top of the plot and most of the blue dots (zone C) are at the bottom, indicating that `LD2` does something to distinguish these two zones. If you thought that zone B was at the bottom of the picture, you might say that `LD2` separates zone A from both of the other two zones. Equally good.

Looking at the group means, the principal difference between zones A and C is in the variable `WidthBF`, which is one of the variables that points (at least a bit) downwards on the biplot, or has one of the larger coefficients on `LD2`. But I wasn't asking for any thoughts of that kind.

(i) (2 marks) Would you say the zones are easy to classify on the basis of the measured variables, or not? Explain briefly.

My answer: I would say the zones are pretty distinct on the plot: greens (B) on the left, reds (A) at the top right, blues (C) at the bottom right. So I would expect the prediction to be good and most of the zones to be gotten right. There are a few exceptions: some reds top left and a blue top right, but it looks as if the zones should be easy to classify.

Any sensible discussion is good, including an opinion that there is too much overlap to classify the groups well (which I don't agree with, but if you can support it I'll accept it).

Was I right? Let's find out, but first I have to read in the data again:

```
library(MASS)
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
summit=read.csv("sumcr.csv",header=T)
summit.s=data.frame(scale(summit[,6:11]),Reach=summit$Reach)
summit.2=lda(Reach~DepthWS+WidthWS+WidthBF+
  HUAreaWS+HUAreaBF+wsgrad,data=summit.s)
pp=predict(summit.2)
table(obs=summit.s$Reach,pred=pp$class)
##    pred
## obs  A  B  C
##   A 16  4  0
##   B  3 43  0
##   C  2  1 19
```

Exactly 10 of the zones got misclassified out of the 88 altogether, which is not very many. This confirms that the zones were pretty different in terms of the measured variables.

I guess I could get a misclassification rate too:

```
data.frame(obs=summit.s$Reach,pred=pp$class) %>%
  mutate(correct=(obs==pred)) %>% group_by(correct) %>%
  summarize(count=n()) %>% mutate(rate=count/sum(count))
## # A tibble: 2 x 3
##   correct count  rate
##   <lgl>   <int> <dbl>
## 1 FALSE      10 0.114
## 2 TRUE       78 0.886
```

The misclassification rate is 11.4%, which is about what you'd expect from the pictures. I leave it to you to decide whether you think that's acceptably low.

In the code, I wanted to divide the `count` for each category (whether or not the prediction was correct) by the total of *all* the counts. This is often what you want to do, so `dplyr` allows you to do this. When you calculate a summary statistic in a `mutate`, it "peels off" a layer of grouping, so that when I calculate `sum(count)` at the end, I don't get the sum of counts for each group (correct or not), as the code suggests I should; instead, the `group_by(correct)` is temporarily undone so that this sum is the sum of *all* the counts. That is, it does what you would like it to do, but if you understand how it works, you can be sure it's doing what you want.

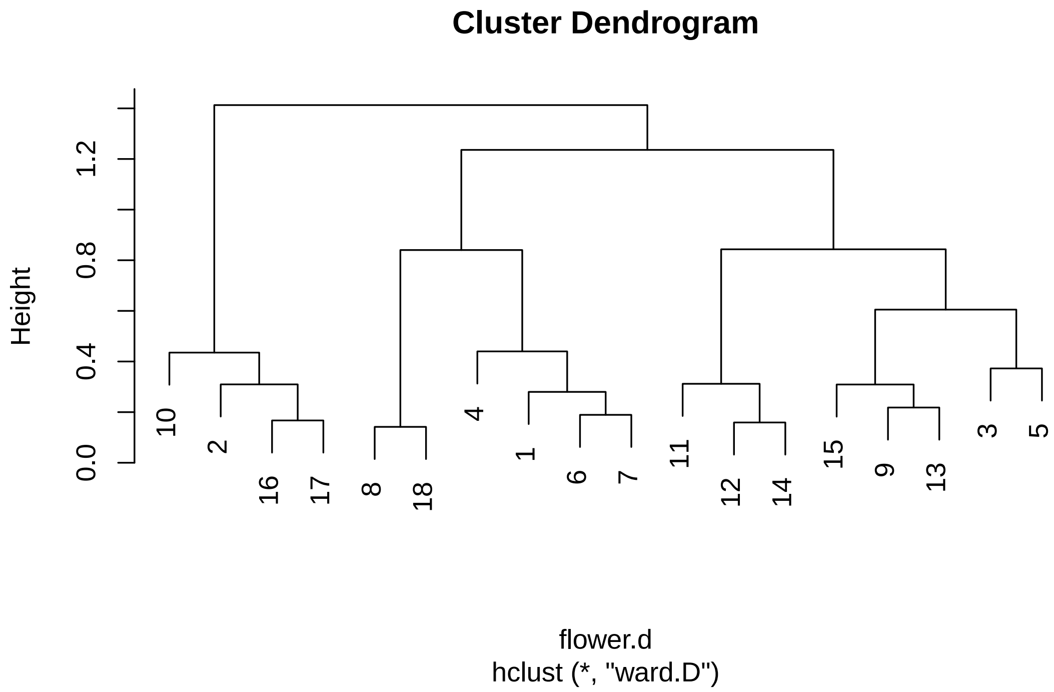
7. Data were collected on 18 popular kinds of flowers, numbered 1–18. For each flower, the following was recorded:

   - `winters`, 1 if the flower may be left in the ground when it freezes, 0 if not.
   - `shadow`, 1 if the flower grows in the shade, 0 if it need to grow in sunshine.
   - `tubers`, 1 if the flower has tubers, 0 if not
   - `colour` (unordered):
     1. white
     2. yellow
     3. pink
     4. red
     5. blue
   - `soil`: grows in dry (1), normal (2) or wet (3) soil. This should be treated as ordered.
   - `preference`: a reviewer's preference ranking, from 1 (best) to 18 (worst). Treat as ordered.
   - `height` in centimetres (a number)
   - `distance` how much space, in centimetres, that should be left between the flowers (a number).

   The data are shown in Figure 20. A `dist` object called `flower.d` is created (code not shown) and used as input into a cluster analysis. The code for the cluster analysis is shown in Figure 21.

   (a) (3 marks) The dendrogram for the cluster analysis in Figure 21 is shown below. How many clusters do you think the flowers should be divided into? Explain briefly. (There is a little space for writing below the dendrogram.) Draw your clusters on the dendrogram, in the same kind of way that `rect.hclust` does, so that it is clear which of the flowers belong in which cluster.

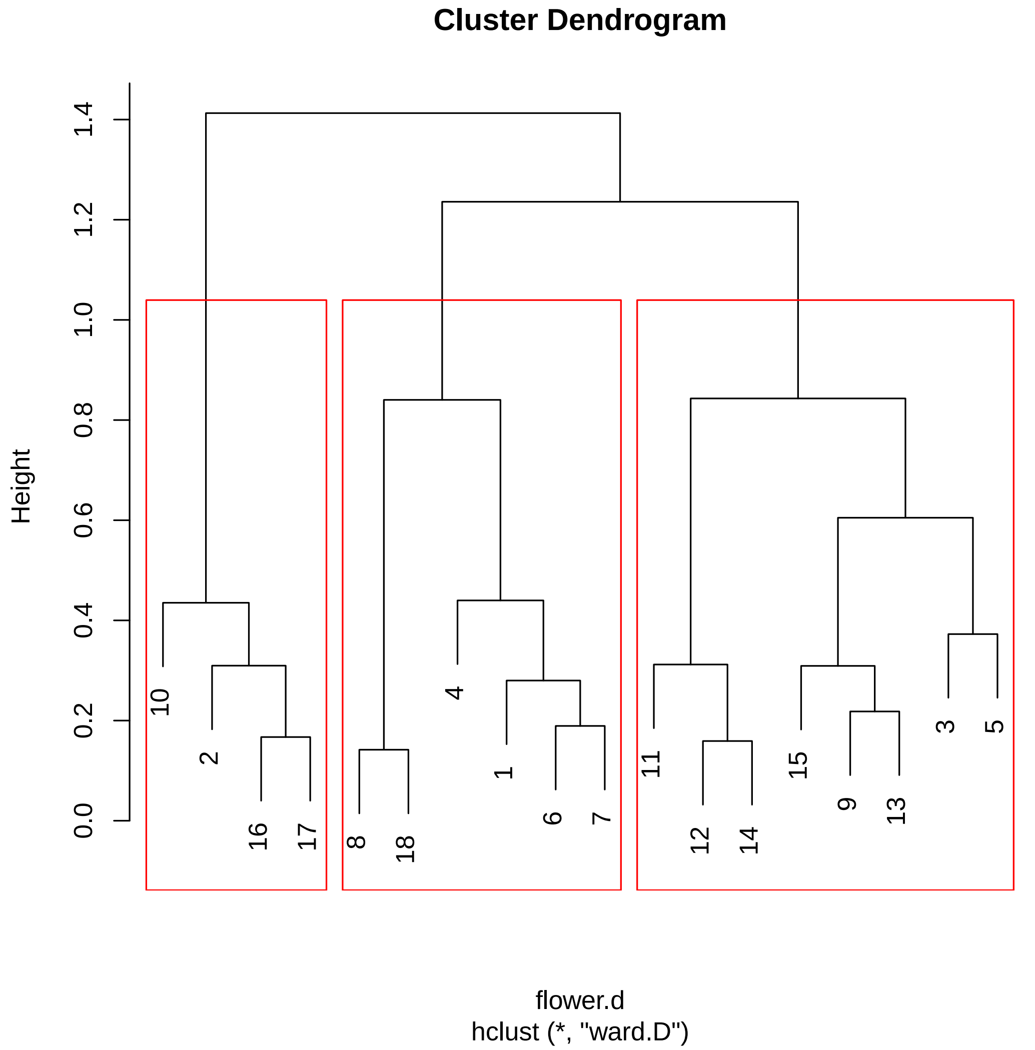
```
## Warning: package 'cluster' was built under R version 3.5.2
```



My answer: My feeling is that three clusters is the best, since the tree doesn't change from height 1.25 all the way down to 0.85 or so. I could also go for anything up to about six clusters. The decision to be made here is that too few clusters will tend to group together dissimilar flowers, while too many will mean splitting up flowers that are similar.

I don't really mind how many clusters you have (but see below), as long as there is *some* justification for it. I'm more interested in whether you can correctly pick out which flowers belong in which of your chosen number of clusters, like this:

```
plot(flower.1)
rect.hclust(flower.1,3)
```



Draw the boxes whatever colour you like, or circle the flowers within each of your clusters, anything like that. As long as I can see which cluster is which. Drawing a line across at the height you "chop the tree" at is smart, since then you can see which flower belongs in which cluster by tracking the tree upwards (or downwards).

The only number of clusters I wouldn't agree with is 4, because it is almost impossible to draw a line across at a height that will get you exactly 4 clusters (and not 3 or 5), because the two splits at height 0.8 happen almost simultaneously. I cannot tell, looking at the dendrogram, which one splits first: is it the 8-18-4-1-6-7 cluster, or the 11-12-14-15-9-13-3-5 one. Cut the tree at a height of, say, 1.0 and get three clusters, or at a height like 0.7 and get 5 (or go down a bit further and get 6, if you like). But your "chop" has to go straight across; if you're going to split the 4-1-6-7 cluster, you must also split 15-9-13-3-5.

(b) (3 marks) Pick two flowers that are in the same one of your clusters (any such pair of flowers will do). By looking back at the original data in Figure 20, does it make sense that the two flowers you chose would be in the same cluster? Explain briefly. (The numbers of the flowers on the dendrogram are the same as the row numbers in the original data frame. You might like to copy the appropriate rows of the data frame `flower0` here to refer to them.)

My answer: Let me pick flowers 16 and 17. (I don't mind what pair *you* pick, as long as they are in the same one of your clusters.)

You can copy the appropriate rows of your data frame, which I think makes it easier to compare:

```
flower0 %>% slice(c(16,17))
```

```
##   winters shadow tubers colour soil preference height distance
## 1       1      0      0      4    2         18    200       60
## 2       1      0      0      2    2         17    150       60
```

Now go along and compare the values one by one, bearing in mind the type of variable each one is. In my case, the factors `winters`, `shadow` and `tubers` all have identical values. The colours of these two flowers are different. The soil type is the same. The preference rank is almost the same, as is the distance. The heights differ by a bit, but these are actually the tallest and the second-tallest flowers, so the heights are not that different. Overall, these flowers, though not identical, are very similar, as they ought to be to be in the same cluster.

Your reasoning will depend on the flowers you chose, but you should find that they are equal on most of the (nominal) factor variables, and close on most of the ordered-factor and numeric variables. This is the point you should try to make. I'd like to see some discussion of all the variables, including any that are different; you ought to be able to make the point that different ones are "close", or there are "only a few" variables showing great difference, or something like that (there is likely to be a little hand-waving, which is fine).

You make life easier on yourself by picking two flowers that are joined together early in the clustering. If you pick two flowers that are joined together later, you might have a hard time demonstrating that they are similar, eg. flowers 4 and 8, which aren't in the same cluster until much later on:

```
flower0 %>% slice(c(4,8))
```

```
##   winters shadow tubers colour soil preference height distance
## 1       0      0      1      4    2         16    125       50
## 2       0      0      1      2    2          7    100       15
```

These are similar on a lot of things, but very different on `preference` and `distance`, so seem to be less similar overall.

(c) (3 marks) The same data underwent a multidimensional scaling, with code shown in Figure 22 and map shown in Figure 23. In part (b) you chose a pair of flowers. Find those same two flowers on Figure 23. (Do not hand in Figure 23.) Are your flowers close together or far apart? Why would you not be so surprised if they were not close together on the map? Explain briefly.

My answer: My pair 16 and 17 are close together on the left side of the map. (My other pair 4 and 8 are a bit further apart, at the bottom of the map.)

Other things being equal, we'd expect flowers in the same cluster to be close together on the map, but the goodness-of-fit in Figure 22 is only 0.54 (the higher of the two figures), which is low, so the map may be a poor representation of the distances in the data (and therefore, actual nearby flowers may not be nearby on the map).

A good intuition here is that two dimensions might not be enough to make a good map of the flowers, and that you might really need three (or more). But the way to tell whether *that* is the case is to look at the GOF. The message from this one is "try more dimensions and see whether it helps".

I didn't tell you anything about how the distances were calculated. I actually used a function called `daisy` from package `cluster`, which I haven't used before (and didn't want to have to explain the details of), but what it does is to make a distance of discrete *and* continuous variables, such as we have here, scaling suitably where necessary: see

`https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/daisy.html`.

There

are a number of variants, all with girls' names (so that `daisy` is really Daisy).

8. Some socioeconomic variables were measured for twelve census districts in Los Angeles (in 1976). The variables were:

   - `district`: a single letter identifying the district
   - `population`: total population
   - `school`: median school years of district residents
   - `employment`: total number of employed people living in district
   - `services`: professional services, whatever that is, in suitable units
   - `housevalue`: median house value in the district (in 1976 dollars).

   We are interested in what makes the districts different, or, equivalently, seeing whether we can find fewer variables to describe them with.

   (a) (2 marks) A principal components analysis is carried in Figure 25, and a scree plot is shown in Figure 26. What is an appropriate number of principal components to use? Explain briefly, using *both* Figures to help you decide.

   > **My answer:** There is a very obvious elbow at 3 on the scree plot, which suggests 2 components. This is further supported in Figure 25, where we see that 2 components explain a pleasant 93% of the variability. So 2 components looks good.
   >
   > For both points, you need to get the elbow from the scree plot, deduce an appropriate number of components, and show that this explains a suitably high fraction of variability from the `summary`. Or start from the `summary`: decide, eg. from the standard deviations (two of them bigger than 1) that two components are good, and then justify it from the scree plot. Either way around is good.

   (b) (2 marks) Component loadings are shown in Figure 27. A blank loading is close to zero and can be ignored. What about a district would make it have a low (negative) score on *component 2*? Explain briefly.

   > **My answer:** Component 2 will have a negative score if `population` and `employment` are large, and `school` and `housevalue` are small.
   >
   > You might be able to make the case that fewer than all four of those variables are important, in which case, you can say whether just those variables need to be high or low. I realize that I messed up the question, since I really wanted you to talk about all four variables, but the way it came out, I couldn't do less than give you full marks if you talked about all the variables *you* thought were important. In retrospect, I would have asked you to name all the variables you thought contributed to component 2 first (or, better, justify why you need all four of the variables with non-trivial loadings), and then the question probably would have been out of 3.

   (c) (4 marks) A plot of the first two principal component scores, with the districts labelled, is shown in Figure 28. Find district D on this plot. Where is it? What should be unusual about this district, given where it is on the plot? Is that the case, looking back at the original data in Figure 24?

   > **My answer:** District D is up at the top of the plot (one easy point). That means it has an unusually positive score on component 2 (one point). This makes it the opposite of the kind of district mentioned in part (b). Thus it has small `population` and `employment`, and large `education` and `housevalue`. (Or, the opposite of whatever you said in (b)). One more point.
   >
   > Let's look back at the original data: district D has one of the smaller populations, a smallish `employment`, almost the largest `education`, and the equal-largest `housevalue`. So this corresponds pretty well to what we'd expect. One last point. (Some kind of concluding comment, saying how well it matches what you'd expect, is a good idea.)

9. A study was made of the classroom behaviour of school students, and factors relating to this. 97 students took part in the study. Over the period of the study, each student's behaviour was classified by the teacher as "deviant" or "non-deviant" (that is, unacceptable or acceptable). In addition, each student was classified as "at risk" or "not at risk" based on their home background, and the adversity level of the student's school was classified as low, medium or high. The primary aim of the study was to investigate the effect of home background and school on behaviour.

   (a) (2 marks) Data of this kind are usually presented as a contingency table, as shown in Figure 30. (The function `ftable` displays the contingency table in the format shown.) Explain briefly why the data frame shown in Figure 29 is more suitable for our analysis.

   My answer: The data in Figure 29 are tidy, with each factor in its own column and all the frequencies in one column. Figure 30 has the factors `behaviour` and `risk` in *rows*, and columns "high", "medium" and "low" that are the *levels* of `adversity`.

   One point for saying that the data frame is tidy (or that the contingency table is not), and one point for saying something about why. You can avoid using the word "tidy", but then you need to say something about why the data frame is suitable (or that the contingency table is not) for the first point.

   A lot of people got 1 point for saying something partially relevant but not demonstrating clear understanding. There was a certain amount of judgement on my part here, but I checked through the pile to be sure I was consistent.

   (b) (2 marks) The first stage of my analysis is shown in Figure 31. My next step is to remove the `behaviour:risk:adversity` term. Why? Explain briefly.

   My answer: In a model with interaction such as this one, the first step is to test the *highest-order* interaction first. If it is not significant, as it is not here (P-value 0.624), I remove it.

   "Because it is not significant" is one point, because it does not explain why I am testing that term in the first place. "Because it has the highest P-value" is *wrong*, because I am not testing any of the other terms until I have decided what to do with this one. I might find you one point for saying this, if you have said something else meaningful.

   (c) (2 marks) The rest of my analysis is shown in Figures 32 and 33. Was my analysis appropriate? Explain briefly why or why not, and if my analysis was not appropriate, what I should have done instead.

   My answer: Follow the analysis through, and make sure I was removing (i) the the least significant of the terms offered by `drop1`, and (ii) I was never removing a term that was significant.

   Everything is good up to and including `classroom.4`. At that point, everything offered for removal is significant, so I *should have stopped there*, keeping the `risk:adversity` interaction. Removing that interaction was a mistake, because it was significant. (Where I stopped, at `classroom.6`, was designed to look sensible, but I should not even have gotten to that point.)

   I saw some really good answers to this one.

   If you want any points here, you need to tell me that I should have stopped at model `classroom.4` (or equivalent, for example that I should not have removed the `risk:adversity` interaction). One point for that, and one for saying why. (The second point should be a gimme if you got the first one.)

   (d) (3 marks) Figure 34 shows some contingency tables for subsets of the data. Based on my analysis, or the correct version of it if my analysis was incorrect, and using the appropriate tables from Figure 34, what do you conclude about the data, especially about any impact of risk and adversity on behaviour? (If you need to see a different subtable to draw your conclusion, explain what subtable you need to see and how you would use it to draw your conclusion.)

My answer: I had to allow for you to go wrong in all kinds of different ways!

The important point here, if you got the previous part right, is to look at tables `tab.4` and `tab.5`, since they are the ones that correspond to the remaining significant terms that cannot be removed. The first one says that most of the students have non-deviant behaviour, *unconnected with any other factors*, while the second one says that students coming from an at-risk home background are more likely to face medium adversity at school, while those students from a not-at-risk home background are more likely to face low adversity at school. You don't have to interpret table `tab.4`, but you need to say somewhere in your answer that neither of the two factors `risk` and `adversity` have a significant effect on behaviour. (None of the interactions involving `behaviour` have remained in the model, which is how you tell.) In summary: one point for describing the relationship between risk and adversity, one point for saying that the majority of students had non-deviant behaviour, and one point for saying what we really care about, which is that neither risk nor adversity have any impact on behaviour.

If you thought my entire analysis of the previous part was correct, then you need to analyze the main effects of `behaviour` and `adversity`, by looking at tables `tab.2` and `tab.3`. These say that more of the students had non-deviant behaviour than deviant, and more of the students faced medium adversity at school than any other level. A point for each of these. To get the third point, you need to make an observation about the other factors being *unconnected* with behaviour, since associations with behaviour were the purpose of the study.

If you thought you needed something else, say what you needed and how it would help you determine the kind of association there is, for example by making up some numbers (although the subtables in Figure 34 could have been gotten by summing up over the missing variables; for example, table `tab.5` could be gotten from table `tab.1` by summing up over deviant and non-deviant behaviour).

Yes, I know that `tab.2` and `tab.4` were the same as each other. I don't know how they both got in there. Maybe I was thinking that depending on your answer to (c), you should be looking at 2&3 or 4&5. But that was definitely a mistake on my part, fortunately not one that had any real impact on the exam.