University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 13, 2022

Aids allowed: my lecture overheads (slides); any notes that you have taken in this course; your marked assignments; my assignment solutions; non-programmable, non-communicating calculator.
This exam has 25 numbered pages of questions.
In addition, you have an additional booklet of Figures to refer to during the exam. Contact an invigilator if you do not have this.
The maximum marks available for each part of each question are shown next to the question part.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.
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.

Question 1 (17 marks)

The data shown in Figure 2 presents the results of an experiment conducted to study the influence of the operating temperature (100C, 125C and 150C) and three faceplate glass types (A, B and C) on the light output of an oscilloscope tube. The data are saved in a dataframe called `gtl`. There are 27 observations in total, three for each combination of glass and temperature.

(a) (2 marks) A plot is shown in Figure 3. The code used to produce the plot is shown above the plot. What do you conclude from the plot?

My answer:

This is an interaction plot: I first calculated the mean light output for each combination of operating temperature and glass type, and then I plotted those against temperature, with the glass type indicated by coloured points and lines.

This means that we should assess whether the lines are parallel. They are clearly not; the trend for glass C goes up and then down as temperature increases, while the trends for the other two glasses go up with temperature. One point. Hence, we would expect to see an interaction between operating temperature and glass type: that is, the effect of temperature on light output will be different for different glass types. The second point.

There is partial credit, likely 1, for other sensible discussion; there might be 1.5 if you get close enough to the idea of interaction without using the word.

(b) (3 marks) What feature of the light output is *not* shown in Figure 3? Describe *in words* a different graph that would show this feature. (Hint: show your understanding by saying what role each variable plays in your graph.)

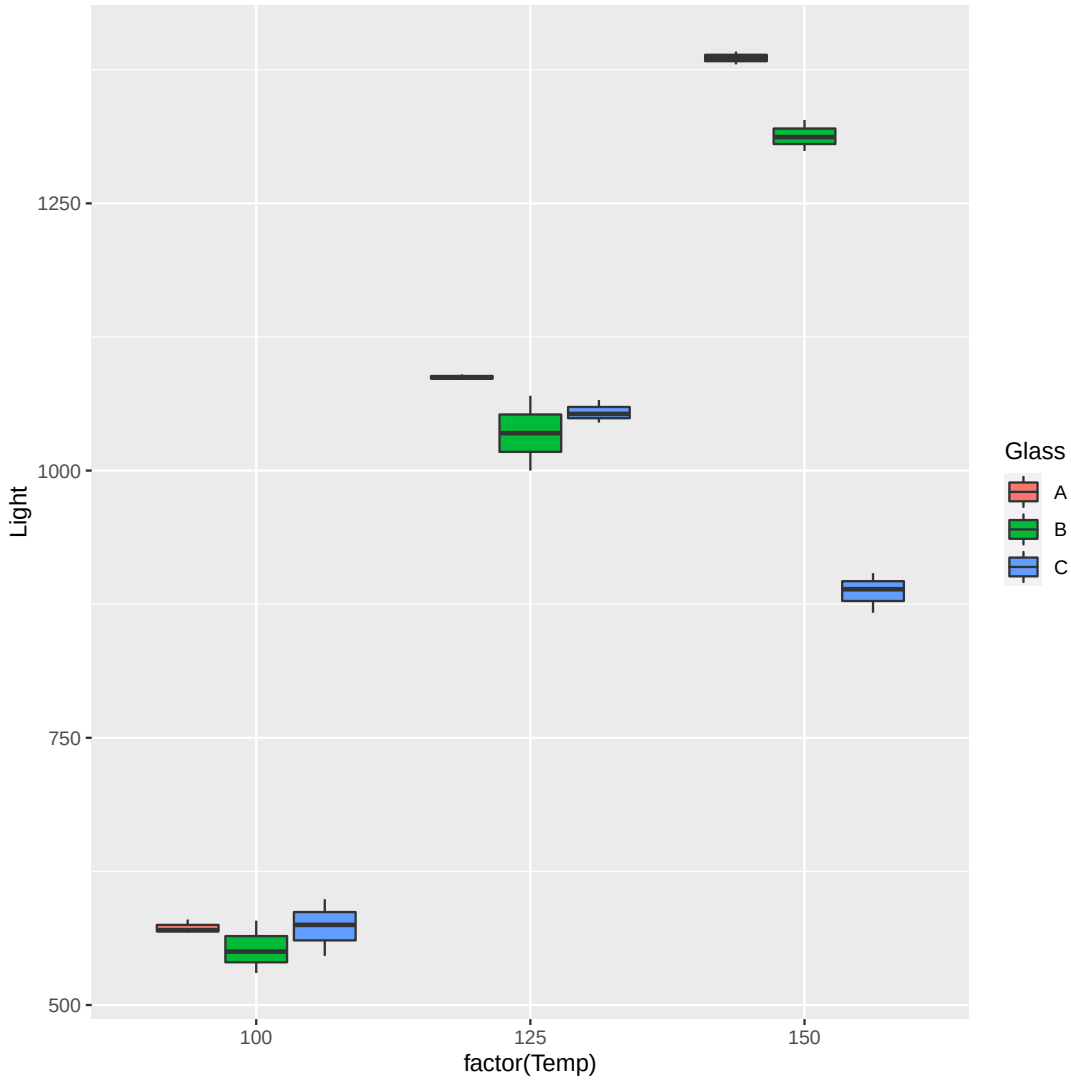
My answer:

The interaction plot does not show the *variability* of light output for each temperature and glass type; it only shows the means. One point. (Alternatively, it doesn't show all the data, which would enable you to see variability/distribution/outliers, which is fine for this point, but then suggests that you're going to talk about a spaghetti plot; see below.)

For the remaining two points: to show the variability as well, we need something like a grouped boxplot, eg. one using temperature as the `x` and glass type as the `fill`.

Extra: this is what such a graph would look like for these data:

```
ggplot(gtl, aes(x = factor(Temp), y = Light, fill = Glass)) + geom_boxplot()
```



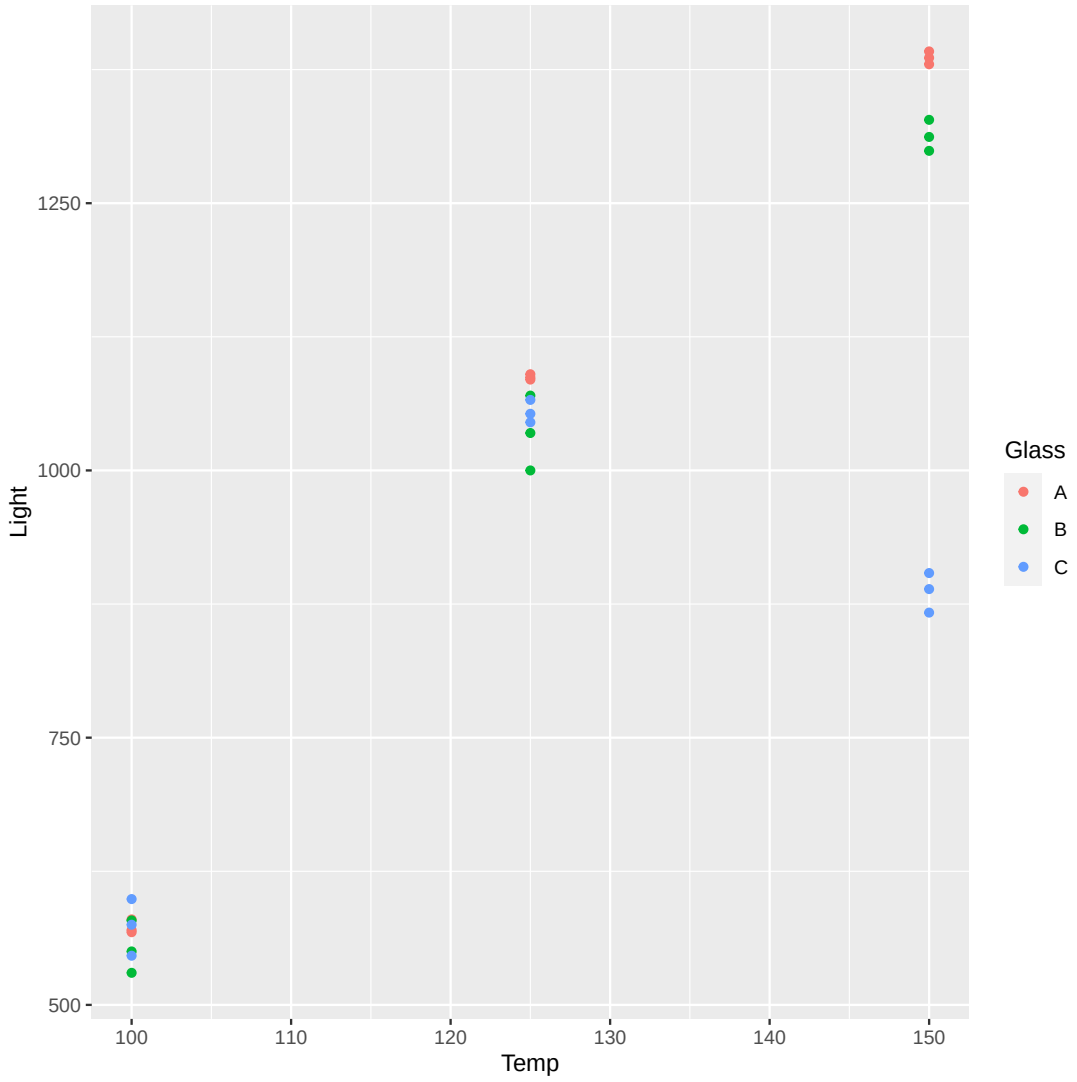
This shows that the variability is very small, and so there is no doubt that the interaction will be significant.

A spaghetti plot is not appropriate here because it is not repeated measures. There are 27 independent observations. If it were repeated measures (over temperature), we would have to have the same experimental units (the exact same pieces of glass) assessed for light output at each of the three different temperatures. (I decided to give this answer one point out of two, because it shows the right kind of intent; it is trying to show all the data in the way that that the scatterplot (below) correctly does for these data.)

If you want to make some other plot, be sure to describe it clearly. For example, if you take

the attitude that temperature is really quantitative (although it is treated as categorical in this question, hence the ANOVA below), you have two quantitative variables (light output and temperature) and one categorical one (glass type) and you could do a scatterplot with the glass types shown by colour:

```
ggplot(gtl, aes(x = Temp, y = Light, colour = Glass)) + geom_point()
```



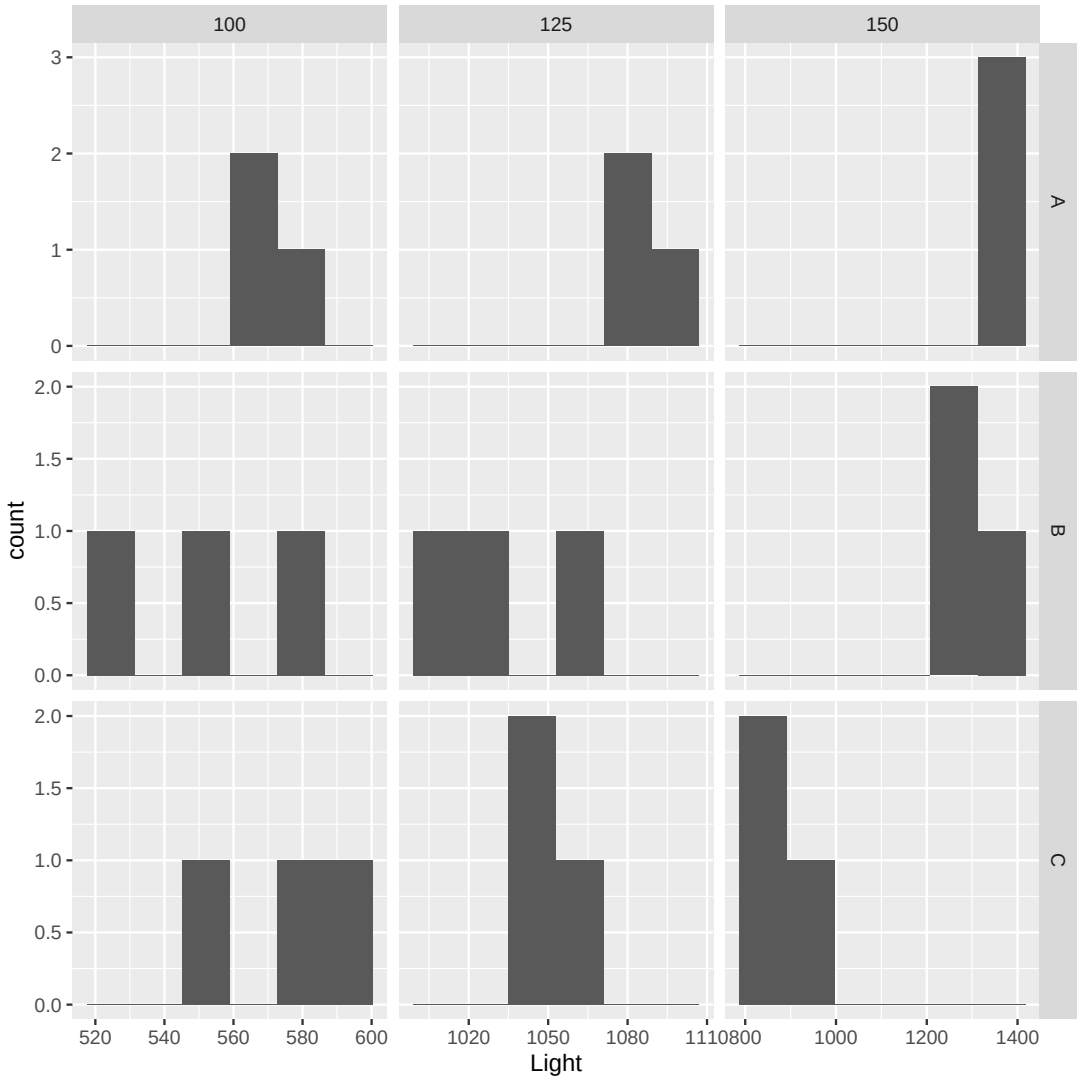
At each temperature, the three dots of each colour are very close together (this shows up especially clearly at temperature 150) and so there is very little variability (to be precise, residual variability once you know glass type and temperature). This is the kind of plot that would precede an analysis of covariance, but if you describe the analysis without describing the

plot, you haven't answered the question I asked. (An analysis of covariance would assume a *linear* effect of temperature on light output, since you're treating temperature as quantitative, while the ANOVA does not: it just says that there is "an" effect of temperature that says the light output could be different for each one in any way at all.)

The scatterplot above is really a lot like a spaghetti plot without the spaghetti strands (lines); there are no points to join by lines because there are no related observations (they are all independent).

Or you could make a histogram of light output for each combination of temperature and glass type, the last two of which you would need to arrange using `facet_grid`. That would come out like this:

```
ggplot(gtl, aes(x = Light)) + geom_histogram(bins=6) +
  facet_grid(Glass ~ factor(Temp), scales = "free")
```



This one is a bit tricky because you need to decide on the number of bins for your histograms, which in turn will depend on whether you have all the histograms on the same scale, or whether (as I have done) you have a different scale for each one. My call is that I wanted to see whether the variability within each combination was large or small, and by having a separate scale for each row and each column, I could eyeball the numbers. Otherwise you need a large number of bins to make each histogram have more than about one bin.

With only three observations per glass-temperature combination, the grouped boxplot is going to be about the best of some not-very-good choices. *After* doing the ANOVA, another option would be to make a boxplot or histogram of the *residuals*; these would all be close to zero

(relative to the scale that the `Light` is measured on), indicating that the variability about the means (for each temperature and glass combination) is small.

(c) (3 marks) An analysis of variance is shown in Figure 4. What do you conclude from it, in the context of the data?

My answer:

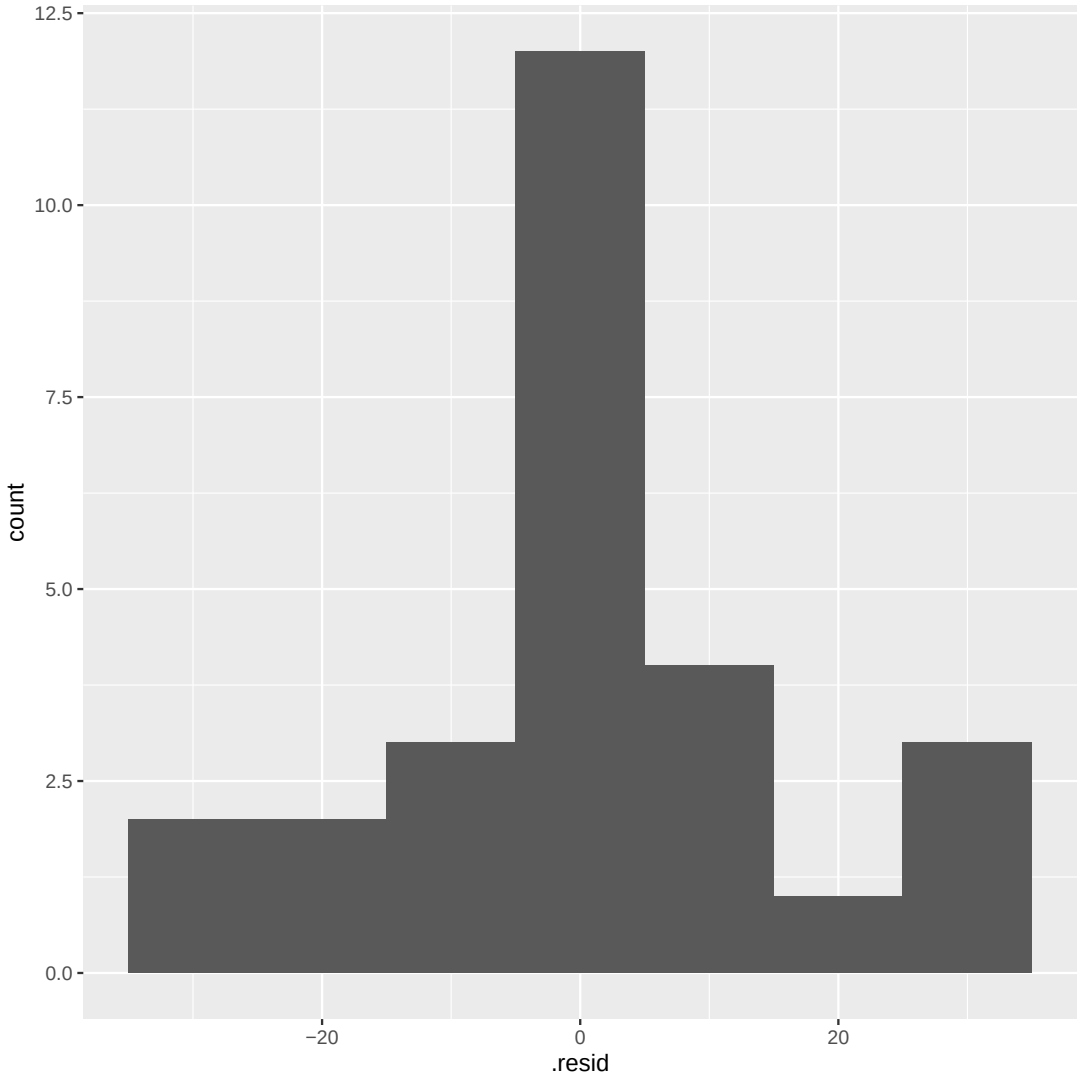
The interaction, with a P-value of 1.25×10^{-14}, is extremely significant. (One point.) This means that the light output depends on the *combination* of temperature and glass type, rather than on either of them individually. One more point.

The third point is for not writing any more! It is a mistake in this kind of analysis to try to interpret the main effects (which are also very significant here) because we need to understand the interaction before we try to interpret anything else. (This is unlike the repeated-measures problem on the midterm because in those, we have no choice but to interpret everything — we cannot remove anything and re-fit — and, in any case, in that one the interaction was only just significant and it was pretty clear from the interaction plot where the main effects were coming from.)

Extra 1: in this one, if you look back at the interaction plot, it is pretty clear that there is there is *not* an effect of temperature that applies to all the glass types, because glass type C behaves differently from the other glass types. That is the reason for looking at simple effects, which we do next.

Extra 2: Now we are in a position to look at residuals. The way to do this is to fit the ANOVA *as a regression* first, and then to look at the residuals as we would in a regression:

```
gtl.2 <- lm(Light~factor(Temp)*Glass, data = gtl)
ggplot(gtl.2, aes(x = .resid)) + geom_histogram(bins = 8)
```



I think this is good *here*, because we were interested in variability, and the `Light` values themselves are much more variable than this:

```
gtl %>% summarize(sd_light = sd(Light))
```

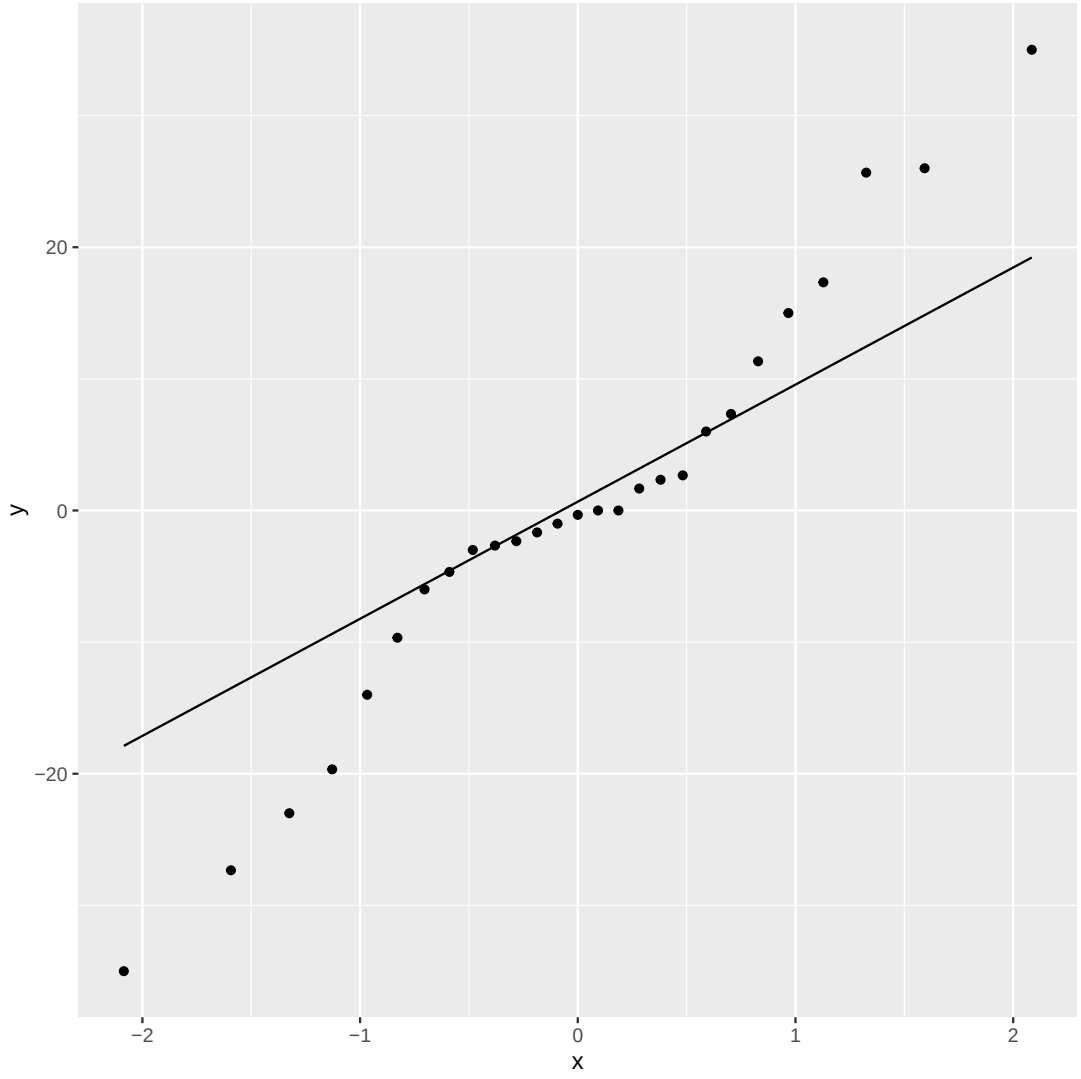
```
##   sd_light
## 1 304.9798
```

That is to say, once you know which temperature and glass type combination you are looking at, there is not much variability left, and even small differences in means can be (and here, are) significant.

The other thing we might be interested in here is the *normality* of the residuals. Our conclusions

are so clear that we don't need to be especially picky about this:

```
ggplot(gtl.2, aes(sample = .resid)) + stat_qq() + stat_qq_line()
```



This is perhaps more long-tailed than we would like, but I don't think there is anything to worry about.

(d) (3 marks) Some more analysis is shown in Figure 5. What is the (statistical) technical term for this kind of analysis? What do you conclude from it, in the context of the data?

My answer:

This is a simple effects analysis (of light output on glass type when temperature is 100). One point.

The F-test is not significant, so at temperature 100, there is no effect of glass type on light output (or, the light output is not significantly different among the three glass types, when temperature is 100). Two points.

You need to say somewhere that this conclusion only applies at temperature 100. The clue is that I showed you the code in which I only used the data for which temperature is 100 for this analysis. (The fact that it is a simple effects analysis ought to clue you in to the idea that we are only looking at one level of *something*.)

(e) (2 marks) Some more analysis is shown in Figure 6. What do you conclude from it, in the context of the data?

My answer:

This one is another simple effects analysis, this time looking at the effect of glass type at temperature 150. There is this time a significant effect of glass type (one point), and the Tukey analysis below shows that *all* the glass types have a different mean light output at this temperature (the second point).

Make sure that you say where you get the conclusion of "all the glass types differ" from. This is *not* justified by the F-test (that only says that there are some differences to be found); it comes from the three significant tests in the Tukey. Also, make sure that you say more than "there are significant differences among glass types"; you have the information to go further than that.

It's not really very useful to compare the sizes of the P-values in the Tukey with each other; they are all significant, so at temperature 150 the light output for all three glass types is different.

(f) (4 marks) Explain briefly how your conclusions from each of the previous two parts are consistent (or are inconsistent) with the graph in Figure 3.

My answer:

This is to say: compare Figure 3 with Figure 5, and then compare Figure 3 with Figure 6.

At temperature 100, there is no significant difference in light output between the three glass types. This shows up on the interaction plot by the three coloured traces being very close together at temperature 100.

At temperature 150, all the glass types have significantly different light output values, and this is consistent with the interaction plot in that all three traces appear far apart. (If you prefer, say that you are surprised that glass types A and B are significantly different because they don't look far apart to you. Either is good. The reason that A and B actually are different is that there is not much variability in actual fact, but you don't know that with the information you have in this exam.)

Two points for each of those.

It is not useful to talk about interactions here, or about increasing trends. What I am looking for is that you know what those simple effects mean and how they show up on the interaction plot.

Question 2 (21 marks)

An Italian wine enthusiast analyzed the chemical composition of 178 wines from one region of Italy. These wines were all of the same type, but the grapes used to make the wines were grown from three different cultivars, labelled 1 through 3 in the data. Wines made from grapes of different cultivars might be expected to have a different chemical composition.

Thirteen different chemicals were measured. In the same order as in the data, these are: Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Nonflavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline. I don't know any more about these variables than what you see here. Part of the dataset is shown in Figure 7.

(a) (2 marks) Figure 8 shows a MANOVA analysis. Why do you think I created the response variable as I did, rather than by using `cbind`? Explain briefly.

My answer:

There are 13 quantitative variables in the data set. `cbind` would require me to list them out one by one, which would be a lot of typing. This way, using `tidyverse` ideas, I don't have to name them at all; they are everything except for `cultivar`, and then I turn them into the required `matrix` at the end.

One point for saying that there are a lot of quantitative variables (or that it is easier this way), and a second point for saying specifically why this way is easier if you have a lot of them.

`cbind` will still work here, but you have to list all the responses you want to include, and it is very easy to miss one. Both `cbind` and this method will create the response matrix, so in your answer you need to contrast the two methods (that is, to say why one is easier, rather than saying what they both do).

I displayed the data using `glimpse` in Figure 7 because there are a lot of variables (so that you could see clearly that there were a lot). It's just this display that makes it look as if the variables are in rows; `wine` is actually an ordinary dataframe with the variables as columns. There is no tidying needed. If I had displayed it the usual way, you might not have noticed how many variables there were, which was the point of the question.

Some thinking outside the proverbial box required here.

(b) (2 marks) What do you conclude from the MANOVA in Figure 8?

My answer:

The P-value is extremely small, so the means on all the variables are not all the same for all three cultivars. Or one or more of the quantitative variables differs for one or more of the cultivars. Or, the cultivar has an effect on the chemical composition of the wine.

We cannot say anything (yet) about which quantitative variables or which cultivars.

I was pretty relaxed about what I accepted here.

(c) (2 marks) A discriminant analysis is shown in Figure 9. Why is it a good idea to do this analysis?

My answer:

Since the MANOVA was significant, one or more of the quantitative variables differs in mean among the cultivars. The discriminant analysis will help us to see which ones distinguish the cultivars.

There are lots of ways to say this. Anything that got at the idea of finding out how the cultivars are different, or how the chemical compositions differ from one cultivar to the next, was good with me. It helps if you show that you know what the groups are here (cultivars) and the response variables (measurements of chemicals).

Hint: copying text out of my notes word for word is a fast way to annoy me. I know what my notes say, because I wrote them. The idea is that you apply those ideas to the data before you, which means using your own words to describe what is going on. I think "will give us some insight", which I saw a few times, are my words too.

Extra: the expression "before you", which is perhaps an odd way of saying "in front of you", reminds me of a pair of Latin anagrams (which, centuries ago, were supposed to reveal a deep truth). This is how you ask "what is truth" in Latin:

Quid est veritas?

If you rearrange the letters in that, you get this Latin sentence:

Est vir qui adest.

which means "It is the man before you". Deep, eh?

OK, I think I'd better get back to marking the exam.

(d) (2 marks) Why are there two linear discriminants in Figure 9? Are they both worth considering? Explain briefly.

My answer:

There are two linear discriminants because there are three groups and 13 variables, and the smaller of 13 and $3 - 1$ is 2.

According to the Proportion of Trace at the bottom of the Figure, both linear discriminants have something to say (69% and 31%) about distinguishing the cultivars (neither proportion is especially small). So it is worth looking at both. (If your opinion is that the second one is small compared to the first one, then say that.)

The proportion of trace for LD1 will *always* be larger than for LD2, because LD1 is (by definition) the one for which the proportion of trace is largest. The issue is whether the proportion of trace for LD2 is large enough to be informative.

(e) (2 marks) In Figure 9, which four of the measured variables make the greatest contribution to the first linear discriminant? Explain briefly.

My answer:

Flavonoids, non-flavonoid phenols, hue, and OD280/315 (of diluted wines).

These have coefficients on LD1 that are farthest from zero (they are actually all negative). An

explanation might help you out if I disagree with your answer.

(f) (2 marks) In Figure 9, which three of the measured variables make the greatest contribution to the second linear discriminant? Explain briefly.

My answer:

Ash, nonflavonoid phenols (again), hue (again). Also the farthest from zero, but here ash is positive and the other two are negative. A discussion of what makes LD2 high (or low) is not needed here, but will help you in the next part (so I was happy to see it here).

If you somehow messed up the previous part but you did this one in what was clearly the same way (that did not make things easier), then I tried to give you full marks for this part. Expect to lose a half point for naming four variables instead of three, or for naming one that was not in the top three.

(g) (2 marks) What kind of values on which variables would make a wine have a very negative score on LD2?

My answer:

A *low* value of Ash (positive coefficient), and high values of nonflavonoid phenols and hue (negative coefficients).

Not negative and positive values for these (they are all positive because they are amounts of something).

(h) (2 marks) A plot of the discriminant scores is shown in Figure 10. Based on this plot, do you think it is difficult or easy to distinguish the cultivars on the basis of their chemical composition? Explain briefly.

My answer:

It should be easy to distinguish them.

It is best to talk about LD1 and LD2 scores: cultivar 1 is low on LD1 and high on LD2; cultivar 2 is middling on LD1 and low on LD2; cultivar 3 is high on LD1 and also high on LD2. This pattern is consistent, so that there should be little chance of mixing the cultivars up. That would be a 2-point answer. Describing where the cultivars are on the page doesn't show as much insight, so 1.5. Just saying that the cultivars are well separated without saying how you know is 1.

(i) (2 marks) In terms of the original variables, what seems to distinguish cultivar 2 from the others? Hint: you have actually already answered this.

My answer:

What distinguishes cultivar 2 from the others on the plot is that it is low (negative) on LD2. But we already know what makes a wine come out low on LD2, since we answered this two parts back: A *low* value of Ash (positive coefficient), and high values of nonflavonoid phenols and hue (negative coefficients). You can check this from Figure 9: cultivar 2 has the lowest mean for Ash, almost the highest for Hue, and the second highest for nonflavonoid phenols. Not perfect, but pretty close.

One point for getting as far as "low on LD2", and the second for repeating (or linking back to) the answer to (g).

(j) (3 marks) Some more output is shown in Figure 11. What is this output, and what do you learn from it? Explain briefly. (Hint: say what the top table is, what the results in it mean, and what the bottom table tells you.)

My answer:

First is a tabulation of each wine's actual cultivar and the cultivar predicted by a discriminant analysis with cross-validation. (The words "cross-validation" need to appear). One point. Minus a half if you miss the cross-validation part.

The table shows that almost all of the wines were classified as the correct cultivar. Only two were misclassified (the ones shown at the bottom): they were both actually cultivar 2, but one of them was misclassified as cultivar 1, and the other one as cultivar 3. Another point for saying something resembling this.

Finally, if you look at the posterior probabilities in the bottom table (the columns labelled `X1` through `X3`), the 2 that was misclassified as a 3 was actually a clear-cut wrong one, but the 2 that was classified as a 1 was a closer call. (Have an opinion about whether it was clear-cut or not; I don't much mind what that opinion is.) The last point for a comment on this.

In general, some kind of sensible discussion will get the point each time, something that shows you know roughly what you're talking about.

Question 3 (14 marks)

ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly) was a multi-site randomized controlled trial conducted from 1999 to 2001. The primary aim of the trial was to test the effects of three distinct cognitive interventions on measures of cognitively demanding daily activities. The three cognitive interventions focused on memory, executive reasoning, and speed of processing. There was also a control intervention, designed to look like a real intervention, but with no actual cognitive effect. Each participant was randomly assigned to one of the interventions. Participants were assessed before training, immediately after training, and after one and two years, for a total of four time points. At each time point, the assessment was done using the Hopkins Verbal Learning Test, on which a lower score indicates better cognitive functioning.

A total of 1,575 older adults completed the trial. Some of the data is shown in Figure 12.

(a) (2 marks) What makes this a repeated measures analysis?

My answer:

Each participant was assessed four times (before training, after training, after 1 and 2 years) rather than just once.

The key point here is the four different times. The three cognitive interventions don't matter here, because each subject only does one of them, and if there weren't repeated measurements at different times, this would be an ordinary one-way ANOVA of the kind that you saw in B27 (or equivalent course).

(b) (2 marks) An interaction plot is shown in Figure 14. Why might you expect not to see an significant interaction between treatment and time for these data? Explain briefly.

My answer:

The traces for the four treatments look approximately parallel, given the amount of variability you might expect to see. (These are humans, so there might be a lot of variability for other reasons.)

Or something equivalent that says that the pattern over time is more or less the same for each treatment, including control.

(c) (2 marks) The repeated measures ANOVA is shown in Figure 15. What feature of this data set do you think led to the interaction term in fact coming out significant?

My answer:

The traces are more or less parallel, so it must have something to do with variability. The summary in Figure 13 indicates that the SDs within each treatment and time are not that small, relative to how much the means differ, so in order for the interaction to come out significant, the sample sizes must have been, and are, very large (which would make the standard errors of the means very small, driving the significance).

One point for saying that the variability is small, which it isn't especially (see Figure 13), but shows the right kind of reasoning. One point also for saying that the control is different from the rest in terms of pattern over time (and was included); I don't think this is enough by itself, but is also sensible reasoning. Saying something like the control group has a different mean from the others is not really enough to get this one point, so half a point.

The question asked for something about the *data*, so telling me that the P-value is small doesn't answer that.

(d) (3 marks) In Figure 15, is there a significant treatment effect? If there is, how would you describe the nature of that treatment effect? (Hint: see Figure 14.)

My answer:

This became a bit more involved since I originally set it: you are supposed to look at the sphericity tests first, except that there is no sphericity test for treatments. The reason behind this is that the treatments don't differ over time (one subject has the same treatment all the way through), and the way the test is done is that the cognitive scores are averaged up over time and then compared, so the sphericity (which is an assessment of how different times compare) is irrelevant for comparing treatments. With that in mind, you can look at the test for treatments in either the univariate or the multivariate tests, and the conclusion will be the same. Hence:

There is a significant treatment effect; look at the P-value, 1.46×10^{-5}, in the MANOVA (univariate or multivariate test, either). One point.

Two points for some sensible discussion of why. Looking at the interaction plot, the control is clearly worst (high is bad), but there may not be much difference between the three real treatments.

I would be wary of asserting that the "speed" treatment is clearly the best since the difference from the other treatments appears small. But, as observed earlier, the sample size is large, so you can use that to support a difference between the real treatments as well. (We actually don't know whether the speed treatment is significantly better than the other real treatments; this would need something like a simple effects analysis to compare the treatments at each time one by one, followed by a Tukey in each case to figure out which differences are actually significant.)

(e) (3 marks) In Figure 15, is there a significant time effect? If there is, how would you describe the nature of that time effect? (Hint: see Figure 14.)

My answer: Again, this has gained some complication since originally set. Now, it would probably be four points altogether, with two marks for getting the right P-value.

First, look at the sphericity test for time. This is strongly significant, P-value 9.2×10^{-17}. That means to look at the Huynh-Feldt adjusted P-value for time, given as 2.4×10^{-101}! (This is very probably the smallest P-value you have ever seen.) Hence, there is a strongly significant time effect (P-value less than 2.2×10^{-16}). One point.

Aside: you might be wondering why the Huynh-Feldt adjusted P-value appears to be *smaller* than the one in the univariate analysis when it is usually larger. But look carefully: the P-value for time in the univariate analysis is actually *less* than 2.2×10^{-16}, and the suggestion here is that it may be quite a lot less than that. End of aside.

This shows up on the interaction plot by the score on each treatment going sharply down after the training and then back up again to about the level before the training, regardless of treatment. Two points for some sensible discussion of how things (consistently) change over time.

Time is treated as categorical in this kind of analysis, so there is no requirement for things to change linearly over time. All that matters is that scores are not the same, somehow, over time, in a way that is at least approximately true for all the treatments. So down and then up again for all the treatments would certainly be a time effect.

(f) (2 marks) Why does it make practical sense that the effect of time would be as you found?

My answer:
When you are training people to do something, in this case to improve cognitive processing, it seems reasonable that right after the participants have learned through the training, that their scores would be better (the second time point). It also seems reasonable that the effect would wear off over time, as participants forget what they have learned, or practice it less. This would explain why the scores get worse between the second and third time points. The improvement seen by the participants in the control group is presumably a placebo effect (they thought they were getting trained in something); by the end, these participants are actually worse than they were at the beginning.

Say something about the reasonableness of immediate improvement, and also about the drop-off over time.

Question 4 (7 marks)

Figure 16 shows the air distances, in miles, between 10 US cities. We are going to see what happens when we perform a hierarchical cluster analysis on these distances.

(a) (4 marks) A complete-linkage cluster analysis is carried out, with the results shown in the dendrogram in Figure 17. Suppose that we want to group the cities into four clusters. Which cities would be in each cluster?

My answer:
"Cut" the dendrogram across somewhere where there are four vertical lines, say at a height of 1000 or a little more, and then follow the vertical lines down to see which cities are in which cluster. From left to right:

- Cluster 1: Seattle, Los Angeles, San Francisco
- Cluster 2: Denver, Houston
- Cluster 3: Miami (only)
- Cluster 4: New York, Washington DC, Atlanta, Chicago

Extra: I originally had you do three clusters, which put Miami in with the four cities in cluster 4, which didn't make much sense to me.

If your clusters are different from mine, I tried to judge whether you made one mistake (3 points) or two (2 points). Putting Miami in with Denver and Houston seemed like an odd reading of the dendrogram. (Three clusters is a good number from the dendrogram, but I asked you for four, so give me four, please.)

(b) (3 marks) Figure 18 shows the latitudes and longitudes of the US cities featured in this question. A city with a more negative longitude is further west, and a city with a more positive latitude is further north. Using this Figure, what do the cities in each of your clusters have in common?

My answer:

If your US geography is good, you can probably do this without Figure 18, but I include the Figure so that you can still make sense of it even if your US geography is not so good:

- Cluster 1: these three cities are the farthest *west* (most negative longitudes). They are actually all on the west coast.
- Cluster 2: these cities are west but not very west, and maybe south-ish.
- Cluster 3: Miami is east and south
- Cluster 4: these cities are east and not-south (least negative longitudes, apart from Miami).

It's a bit of a struggle to call the cities in cluster 4 "north" when one of them is Atlanta, but "mainly north-east" would be OK as a description.

If your clusters differed from mine, you get 3 points here for an honest effort to say what your clusters have in common location-wise. Since I gave you the longitudes and latitudes, I was looking for something that said which part of the US the cities in each cluster were, like "west" or "south-east" (eg. Miami, the farthest south and one of the farthest east).

Extra: you might immediately think of calling Denver and Houston "mid-west", but that term typically applies to an area in the *northern* US, starting at about Cleveland (a bit east of Chicago) and going west from there about halfway to the west coast. Places like Chicago, Milwaukee, and Minneapolis are usually thought of as part of the midwest, and somewhere like Fargo, North Dakota. Think of places where it gets really cold in winter.

Question 5 (12 marks)

A basketball fan collected stats on around a thousand past and present National Basketball Association (NBA) players. Some information about basketball in general and the stats listed below is given in Figure 19. We will consider the following stats:

- `fg_pct`: "field goal percent": the number of (2-point) shots made, divided by the number attempted
- `3fg_pct`: the number of 3-point shots made, divided by the number attempted
- `ft_pct`: the number of free throws made, divided by the number attempted
- `oreb`: the number of offensive rebounds per game
- `dreb`: the number of defensive rebounds per game
- `ast`: the number of assists per game
- `stl`: the number of steals per game
- `blk`: the number of shots blocked per game
- `tov`: the number of turnovers per game

The first three, labelled `pct`, are actually proportions, between zero and one; the others are made comparable for different players by dividing the total number made by the number of games that player played. (Some players have played many more games than others.)

Player stats in basketball generally fall into one of three categories: shooting effectiveness (the ones that end with `pct` in this question), rebounding, and the others (that benefit or hurt the team that the player belongs to).

Some of the data is shown in Figure 20, and a scree plot is shown in Figure 21.

(a) (2 marks) How many factors should be used for a factor analysis? Explain briefly.

My answer:

Use the scree plot in Figure 21. There is a big elbow at 3, so we should use two factors. I would also be happy with a point of view that this is too far up the mountain, and that therefore we should consider the elbow at 5 and use 4 factors. (The consideration is the same as for principal components.)

We have a lot of data, over 1000 players, but there are only 9 variables, and so a number of factors (substantially) less than 9 is called for here. You could make a case for 4 here, but more than that is not really offering insight.

One point for finding an elbow, and one for subtracting one to get a number of factors. Make sure you "connect the dots": say where the elbow is, and then say how many factors you are going to use.

I said in the question that player stats could be classified as shooting, rebounding, and the rest, but the issue here, if you think there should therefore be three factors, is whether that is in any way justified by the scree plot (it seems that it is not).

(b) (4 marks) A factor analysis is shown in Figure 22. This uses two factors, which may or may not be the same as your preferred number of factors from the previous part. Which of the original variables are an important part of each of the two factors? Explain briefly.

Question 5 continues... This page: 6 marks.

My answer:

In each case, we are looking for loadings far from zero (either positive or negative):

- for the first factor, the two rebounding stats `oreb` and `dreb`, also blocks and field goal percentage, and optionally 3-point shooting and free-throw shooting, with these last two being negative. (That means, someone who scores high on factor 1 will be *low* on these last two.)
- on factor two, assists, steals, and turnovers.

I have no objection if you take fewer variables, as long as they are the ones with the largest loadings in size. I didn't say how many variables to take (maybe I should have done). Draw the line for a "large" loading wherever you see fit. This is factor analysis, with a varimax rotation, so the loadings ought to divide themselves into large and small ones, but this is clearer on factor 2 than it is on factor 1.

It is nice if you actually say what the variable names actually represent. I wasn't picky about that here, but I will probably be more so in the next part.

Someone who scores high on factor 1 is good at the "in-the-paint" parts of the game (close to the basket): rebounding, 2-point shooting (close to the basket), blocking shots: things that involve competing with other players. Someone who scores low on factor 1 is bad at these and good at three-point and free-throw shooting where the emphasis is on pure skill. Someone who scores high on factor 2 will have a lot of assists, steals, and turnovers, which are characteristic of someone who handles the ball a lot (like Magic Johnson, as we see below).

Extra 1: something coming out of this that is unusual in most sports is that there is no real distinction between good attacking players and good defensive players. In factor 1, a player that scores high is good at (2-point) shooting and taking offensive rebounds, but the same player is also likely to be good at the other end of the court as well: taking defensive rebounds and blocking shots.

Extra 2: the output also shows that these two factors explain 61% of the total variability, which you might find low for something like this. (Four factors would explain 85% of the variability, which you might in that case find more satisfactory. If I didn't want to get the grading done, I would definitely explore this.)

(c) (6 marks) A plot of factor scores of all 1,000 players would be very hard to read, so we focus on just three players. A plot of the factor scores for the three players is shown in Figure 23. In addition, the original data for these three players are shown in Figure 24, and the percentile ranks on all the variables for those data are shown in Figure 25. For each of the three players, based on their positions on the plot, which of the original variables would you expect to be high or low, and is that indeed the case? Explain briefly.

My answer:

To take the three players in turn, two points each:

- Magic Johnson is easily the highest on factor 2. This means that we would expect him to be high on assists, steals, and turnovers. He is the highest of all on assists (100th percentile), and also very high on steals (98th percentile) and almost the highest on turnovers. (Extra: Johnson's position, point guard, meant that he handled and passed the ball a lot, and players in this position often have a lot of turnovers. This means that he was far from being a bad player; indeed, he was one of the all-time greats.)
- Dennis Rodman is high on factor 1. This means high on rebounding, field goal percentage, and blocks, and possibly also low on three-point shooting and free-throw shooting. Rodman's rebounding (offensively and defensively) is almost the best in the entire dataset, and his field goal percentage is in the 92nd percentile. He is only at the 69th percentile on blocks, but the overall picture is high in the right places. Also, his three-point shooting is not good (37th percentile) and his free-throw shooting is awful (5th percentile)! He was a notoriously bad free-throw shooter.
- Steve Novak is low on factor 1, and also low on factor 2. That means he should be opposite of both Rodman and Johnson, basically low on everything except for three-point and free-throw shooting. He is above the 98th percentile on those, and near the bottom on pretty much everything else, entirely as expected.

Assess the variables that you thought were important in each factor, whatever they were, or some reasonable subset of them. First say what variables should be high (or low), and then assess whether those variables actually were (using the percentile ranks). Say what the variables are in English, rather than using the abbreviated variable names, on the basis that somebody (me!) wants to read and understand what you have to say.

Strictly speaking, you should refer to Figure 25 to assess how these players rank relative to the entire dataset, and should show that you know what percentile ranks mean (as I did above).

Extra: with the ones that are "percentages", don't confuse the percentile ranks with the original data. For example, Dennis Rodman was at the 5th percentile for free-throw shooting, but that doesn't mean he only shot 5% of his free throws (that would be truly awful); he actually shot 58% of them (Figure 24), but almost all the players in the dataset did better than that. Evidently Rodman's value was "under the basket" at both ends of the court: shooting, mainly from close to the basket, and especially rebounding, at which he was a master.

Question 6 (11 marks)

The Scouts is an organization that provides outdoor activities and leadership opportunities for boys aged 11–14. It is believed that boys enrolled in Scouts are less likely to be involved in criminal activities. In

a survey, 800 boys of Scouting age were classified by socioeconomic status, whether or not they were enrolled in Scouts, and whether or not they were classified as a juvenile delinquent. Juvenile delinquency is defined as "the act of participating in unlawful behavior as a minor or individual younger than the statutory age of majority". The data are shown in Figure 26.

(a) (2 marks) Two tables are shown in Figure 27. One of the numbers in the (bottom) table is 0.0877. What does that number tell you?

My answer:

This says that 8.77% of the boys enrolled in Scouts are juvenile delinquents. (Or, "out of the boys enrolled in Scouts, 8.7% of them are ..." or something equivalent.)

Be careful: this is a conditional proportion — you know the boys in this row are Scouts, so it is the proportion of boys that are delinquent *given* that they are Scouts. Specifically, it is this:

```
33/(33+343)
```

```
## [1] 0.08776596
```

It is not the joint proportion of boys that are both Scouts and delinquent (out of all the boys), which would be this:

```
33/(360+64+343+33)
```

```
## [1] 0.04125
```

Using the word "of" ("of Scouts") is likely to get you to the right answer; using the word "and" ("is a Scout and is delinquent") is likely to get you to the wrong one.

Therefore, be careful how you say it.

(b) (2 marks) In Figure 27, what is the purpose of the `margin = 1` in the code to make the second table? (This is connected with your answer to the previous part; if you answer this part elsewhere, you get credit for this part too.)

My answer:
This makes the *rows* add up to 1, so that it makes sense to talk about "out of the boys enrolled in Scouts". In this context, being a Scout is the explanatory variable, and we want to see whether that has any impact on a boy being a juvenile delinquent.

(c) (2 marks) What is your overall conclusion from Figure 27, in the context of the data?

My answer:
Fewer of the boys who are in Scouts are juvenile delinquents, approximately 9% compared to 15%. If you got to a conclusion like this, you should have two points. Something less insightful, like saying the majority of boys are not juvenile delinquents whether or not they are Scouts (true, but not very illuminating) is probably only one.
Keep that in mind, because we are going to change our mind about this conclusion shortly.

(d) (3 marks) Some analysis is shown in Figure 28. What do you conclude from it, and how is that different from what you concluded earlier? Explain briefly.

My answer:
At the end (which is the important thing), we have remaining associations between (i) socioeconomic status and whether or not the boy is in Scouts, and (ii) socioeconomic status and whether or not the boy is a juvenile delinquent. That's two points. For the third, say something about what is *not* there and why: there is actually *no* association between between scouting and juvenile delinquency, because it was removed from the model, and what has actually happened is that the link that we thought we saw is entirely because of socioeconomic status. We'll see exactly what that link is in the next part.

(e) (2 marks) Some more tables are shown in Figure 29. Using these tables, how can you add to your conclusion from the previous part?

My answer:
These say (i) that boys from families with high socioeconomic status are more likely to be Scouts (and those with lower status are less likely), and (ii) boys from families with high socioeconomic status are less likely to be juvenile delinquents (and those with low are more likely).
In other words, it seems that socioeconomic status is driving everything. Strictly speaking, we only have associations here, but it seems likely that the causal effects are this way around. For example, it's hard to imagine the fact of a boy being in Scouts being the cause of his family's high socioeconomic status. It makes much more sense the other way around.
Extra: socioeconomic status is playing the role of a "lurking variable" here: it is actually the reason for an apparent association between scouting and juvenile delinquency, an association that goes away when you look at socioeconomic status. Causal analysis people call a variable

like our socioeconomic status a "confounder" because it confounds or confuses the association between scouting and juvenile delinquency: we don't know what the real relationship is, or indeed whether there is one at all, without doing further analysis like the log-linear model, because the socioeconomic status is getting in the way.

There were a lot of students who did a nice job of piecing all this together.

Use this page if you need more space to write your answers. Be sure to label any answers here with the question and part that they belong to.

End of Exam This page: 0 marks.