Mixed assessment 1: Mark scheme

Marking scheme for student ID: 1004153364

Question 1

Academic integrity question.

You selected: "In submitting this assessment, I confirm that my conduct during this assessment adheres to the Code of Behaviour on Academic Matters. I confirm that I did NOT act in such a way that would constitute cheating, misrepresentation, or unfairness, including but not limited to, using unauthorized aids and assistance, impersonating another person, and committing plagiarism."

Thank you for confirming that your conduct adhered to the Code of Behaviour on Academic Matters..

Question 2 (1 point)

Question 2, field 1

What is the name of your ship? Hint: check out the object `ship_name`. (1/3 point)

```
ship_name
```

```
## [1] "SS Breustreats"
```

Your answer: SS Breustreats.
Correct answer: SS Breustreats.

Well done, 1/3 points.

Question 2, field 2

What is the name of the Communications Officer?

```
# ANSWER
name_comms <- crew_data %>%
  filter(position == "Communications Officer") %>%
  distinct(name) %>%
  as.character()
```

Your answer: Crystal Hernandez.
Correct answer: Crystal Hernandez.

Well done, 1/3 points.

Question 2, field 3

How many crewmembers are in this dataset?

```
# ANSWER
n_crew <- crew_data %>%
  distinct(crew_id) %>%  #note use of crew_id instead of name
  nrow()
```

Your answer: 251.
Correct answer: 251.

Well done, 1/3 points.

Question 3 (1 point)

Prep from untimed component (task 2, part 1)

The Records Officer lets your know that there is a typo in the crew dataset, where 'Engineering' has been misspelled somewhere, (maybe in one of the position titles?) but unfortunately they can't remember where or how. Find the mistake, fix it (and save that fix in the original `crew_data`) and then calculate what proportion of people in the Engineering subdivision have 'engineer' or 'engineering' in their position title.

```
# ANSWER

## fix the typo
crew_data <- crew_data %>%
   mutate(position = str_replace(position, "Enigneering", "Engineering"))

## calcualte the percentage
perc <- crew_data %>%
  distinct(crew_id, .keep_all =TRUE) %>%
  filter(sub_division == "Engineering") %>%
  mutate(eng = str_detect(position, "Engin")) %>%
  summarise(prop = mean(eng)) %>%
  as.numeric() %>%
  round(.,2)

# or equivalently

crew_data %>%
  filter(sub_division == "Engineering") %>%
  mutate(eng = case_when(str_detect(position, "Engin") ~ 1,
                         TRUE ~ 0)) %>%
  summarise(prop = mean(eng)) %>%
  as.numeric() %>%
  round(.,2)
```

```
## [1] 0.62
```

Timed assessment question

What proportion of crewmembers in the engineering sub-division have 'engineer' or 'engineering' in their position title? Round to two decimal points, e.g., 0.24 or 0.99 etc.

Your answer: 0.62.
Correct answer: 0.62.

Well done, 1 point.

Prep from untimed component (task 2, part 2)

Create a new variable in `crew_data` called `full_team` that indicates both the duty shift and the team each person is assigned to.

- You may find the `str_c()` function useful.
- You can specify how the values you're sticking together are separated with the `sep` parameter, e.g., `str_c(var1, var2, sep = " ")` would put a space between the values of var1 and var2 when sticking them together.
- Don't forget that `mutate()` helps you make new variables.

```
# ANSWER
crew_data <- crew_data %>%
  mutate(full_team = str_c(duty_shift, shift_team, sep = " "))
```

Prep from untimed component (task 3, part 1)

Create a new dataset called `week1` that filters to only the observations for week 1. You must also reverse the levels of the `duty_shift` factor in `week1` so that the order is: Gamma, Delta, Beta, Alpha. You can test if you've achieved this by running `table(week1$duty_shift)`. The table should be ordered with Gamma first.

```
# ANSWER
week1 <- crew_data %>%
  filter(week == 1) %>%
  mutate(duty_shift = fct_rev(duty_shift))

table(week1$duty_shift)
```

```
##
## Gamma Delta  Beta Alpha
##    49    49    72    81
```

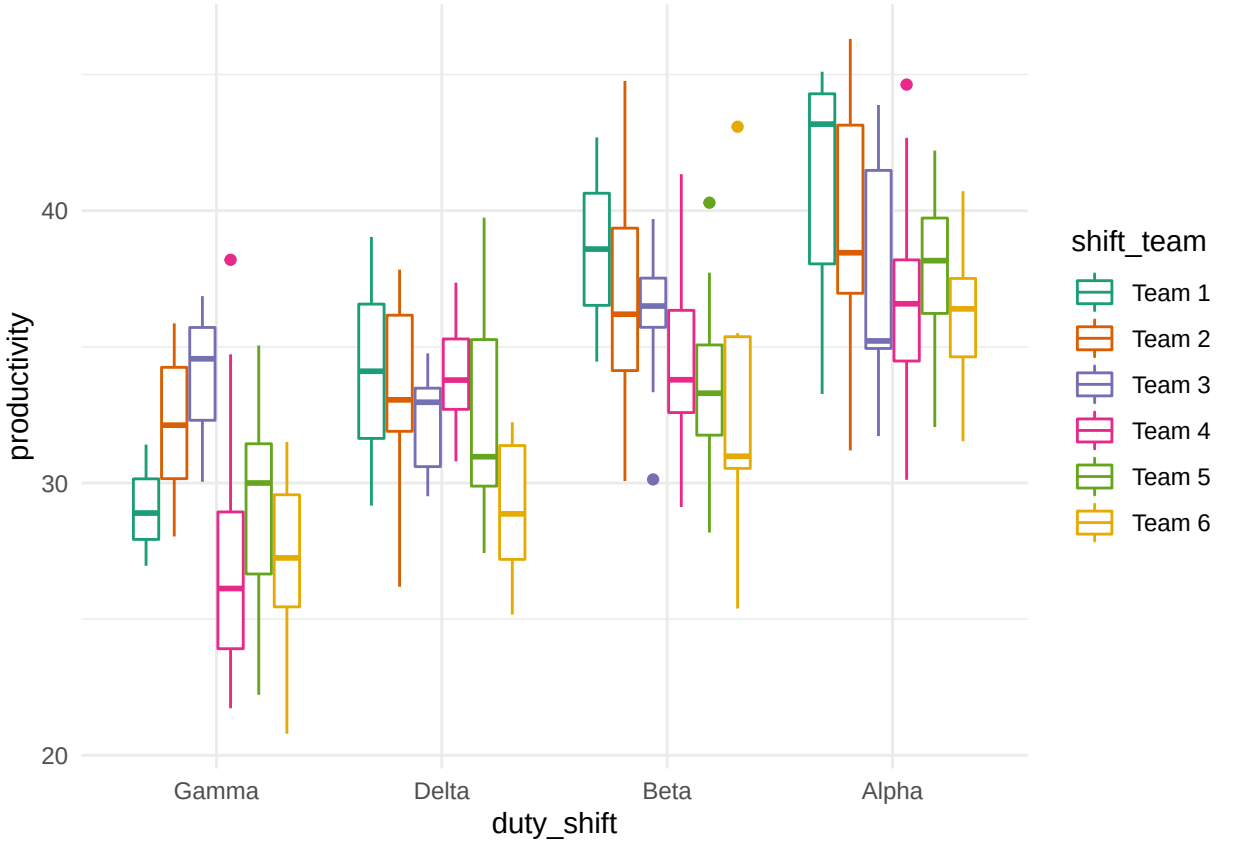
Question 4 (2 points)

Prep from untimed component (task 3, part 2)

Using the `week1` dataset you created, create a plot with `productivity` on the y-axis, `duty_shift` on the x-axis and coloured boxplots for each `shift_team`. Use the "Dark2" colour palette from colour brewer.

- geom_boxplot() is the geometry that creates boxplots.
- use the colour aesthetic to get different boxplots for each `shift_team`
- `scale_colour_brewer()` will allow you to choose the Dark2 palette (when completed appropriately).

```r
# ANSWER
week1 %>%
  ggplot(aes(x = duty_shift, y = productivity, colour = shift_team)) +
  geom_boxplot() +
  scale_colour_brewer(palette = "Dark2") + # following the palette instruction is important
  theme_minimal()
```



Timed assessment question

Which duty shift and shift team combination represents the yellow boxplot with the highest median?

Your answer: Alpha shift and Team 6.

Correct answer: Alpha shift and Team 6.

Shift is correct.

Team is correct.

2 points.

Question 5 (2 points)

Which ONE of the following statements is most appropriate with respect to w1_shift?

You should not need to run any additional code or tests to answer this.

Your answer: There is no reason to believe any of our linear regression assumptions are violated so it seems okay to proceed.

Correct answer: "There is no reason to believe any of our linear regression assumptions are violated so it seems okay to proceed."

Why is this correct? In this model, there is only one observation per crew member, so no independence issue due to repeated measures. It does seem believable that there could be a straight-line relationship between weeks since shore leave and productivity. It does seem believable that the response variable will be normally distributed. It is **not** true that linear regression is robust to all violations of its assumptions.

Your score: 2 points.

Question 6 (1 point)

Prep from untimed component (task 3, part 3)

Using the `week1` data, fit a linear model called `w1_shift` where `productivity` is the response and `duty_shift` is the only predictor. Run `summary` and `confint` on the model.

```
# ANSWER
w1_shift <- lm(productivity ~ duty_shift, data = week1)
summary(w1_shift)
```

```
##
## Call:
## lm(formula = productivity ~ duty_shift, data = week1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6693 -2.8446 -0.3136  2.5360  9.7028
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      28.9236     0.5592  51.720  < 2e-16 ***
## duty_shiftDelta   3.6396     0.7909   4.602  6.7e-06 ***
## duty_shiftBeta    6.1365     0.7250   8.465  2.3e-15 ***
## duty_shiftAlpha   9.0751     0.7085  12.809  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.915 on 247 degrees of freedom
## Multiple R-squared:  0.4176, Adjusted R-squared:  0.4105
## F-statistic: 59.03 on 3 and 247 DF,  p-value: < 2.2e-16
```

```
confint(w1_shift)
```

```
##                     2.5 %    97.5 %
## (Intercept)     27.822085 30.025027
## duty_shiftDelta  2.081842  5.197272
## duty_shiftBeta   4.708620  7.564431
## duty_shiftAlpha  7.679708 10.470530
```

Prep from untimed component (task 3, part 4)

Fit three additional linear models and run summaries on them:

- Name the first model `w1_team`. It should have `productivity` as the response and then `shift_team` as the only predictor. `week1` is still the data to use.
- Name the first model `w1_int`. It should have `productivity` as the response and then the main effects and interaction of `duty_shift` and `shift_team` as the predictors. `week1` is still the data to use.
- Name the second model `w1_full`. It should have `productivity` as the response and `full_team` as the only predictor. `week1` is still the data to use.

```
# ANSWER
w1_team <- lm(productivity ~ shift_team, data = week1)
summary(w1_team)
```

```
##
## Call:
## lm(formula = productivity ~ shift_team, data = week1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0732  -3.2955  -0.0553   3.2835  12.0758
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        37.1631     1.2397  29.978  < 2e-16 ***
## shift_teamTeam 2  -0.8656     1.3908  -0.622   0.5343
## shift_teamTeam 3  -1.6469     1.5567  -1.058   0.2911
## shift_teamTeam 4  -3.3604     1.3959  -2.407   0.0168 *
## shift_teamTeam 5  -3.3121     1.3959  -2.373   0.0184 *
## shift_teamTeam 6  -6.1545     1.4536  -4.234 3.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.801 on 245 degrees of freedom
## Multiple R-squared:  0.131,  Adjusted R-squared:  0.1132
## F-statistic: 7.385 on 5 and 245 DF,  p-value: 1.807e-06
```

```
w1_int <- lm(productivity ~ duty_shift*shift_team, data = week1)
summary(w1_int)
```

```
##
## Call:
## lm(formula = productivity ~ duty_shift * shift_team, data = week1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9293 -2.3868 -0.1157  2.4744 10.7239
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      29.08851    2.13577  13.620  < 2e-16 ***
```

```
## duty_shiftDelta                   5.01534    3.37695   1.485   0.1389
## duty_shiftBeta                    9.49154    3.02043   3.142   0.0019 **
## duty_shiftAlpha                  11.80186    2.55273   4.623 6.34e-06 ***
## shift_teamTeam 2                  3.03130    2.55273   1.187   0.2363
## shift_teamTeam 3                  4.73452    3.02043   1.567   0.1184
## shift_teamTeam 4                 -1.61534    2.40947  -0.670   0.5033
## shift_teamTeam 5                  0.09424    2.40947   0.039   0.9688
## shift_teamTeam 6                 -1.91238    2.35350  -0.813   0.4173
## duty_shiftDelta:shift_teamTeam 2 -3.67287    3.79623  -0.968   0.3343
## duty_shiftBeta:shift_teamTeam 2  -4.78379    3.44067  -1.390   0.1658
## duty_shiftAlpha:shift_teamTeam 2 -4.79613    3.02582  -1.585   0.1143
## duty_shiftDelta:shift_teamTeam 3 -6.57346    4.32460  -1.520   0.1299
## duty_shiftBeta:shift_teamTeam 3  -7.18594    3.87982  -1.852   0.0653 .
## duty_shiftAlpha:shift_teamTeam 3 -8.20742    3.57610  -2.295   0.0226 *
## duty_shiftDelta:shift_teamTeam 4  1.47979    3.74383   0.395   0.6930
## duty_shiftBeta:shift_teamTeam 4  -2.59321    3.34246  -0.776   0.4387
## duty_shiftAlpha:shift_teamTeam 4 -2.23369    2.91903  -0.765   0.4449
## duty_shiftDelta:shift_teamTeam 5 -1.48722    3.72718  -0.399   0.6903
## duty_shiftBeta:shift_teamTeam 5  -5.21479    3.34246  -1.560   0.1201
## duty_shiftAlpha:shift_teamTeam 5 -2.98389    2.92668  -1.020   0.3090
## duty_shiftDelta:shift_teamTeam 6 -3.24796    3.75390  -0.865   0.3878
## duty_shiftBeta:shift_teamTeam 6  -3.62815    3.47209  -1.045   0.2972
## duty_shiftAlpha:shift_teamTeam 6 -2.88222    2.95600  -0.975   0.3306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.699 on 227 degrees of freedom
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.4736
## F-statistic: 10.78 on 23 and 227 DF,  p-value: < 2.2e-16
```

```
w1_full <- lm(productivity ~ full_team, data = week1)
summary(w1_full)
```

```
##
## Call:
## lm(formula = productivity ~ full_team, data = week1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9293 -2.3868 -0.1157  2.4744 10.7239
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               40.890      1.398  29.245  < 2e-16 ***
## full_teamAlpha Team 2     -1.765      1.625  -1.086 0.278475
## full_teamAlpha Team 3     -3.473      1.915  -1.814 0.071005 .
## full_teamAlpha Team 4     -3.849      1.648  -2.336 0.020369 *
## full_teamAlpha Team 5     -2.890      1.661  -1.739 0.083321 .
## full_teamAlpha Team 6     -4.795      1.789  -2.681 0.007886 **
## full_teamBeta Team 1      -2.310      2.553  -0.905 0.366404
## full_teamBeta Team 2      -4.063      1.648  -2.466 0.014419 *
## full_teamBeta Team 3      -4.762      1.823  -2.612 0.009602 **
## full_teamBeta Team 4      -6.519      1.661  -3.924 0.000115 ***
## full_teamBeta Team 5      -7.431      1.661  -4.473 1.22e-05 ***
```

```
## full_teamBeta Team 6    -7.851      1.977  -3.970 9.63e-05 ***
## full_teamDelta Team 1   -6.787      2.966  -2.288 0.023052 *
## full_teamDelta Team 2   -7.428      1.734  -4.283 2.72e-05 ***
## full_teamDelta Team 3   -8.625      2.166  -3.982 9.20e-05 ***
## full_teamDelta Team 4   -6.922      1.823  -3.797 0.000188 ***
## full_teamDelta Team 5   -8.179      1.789  -4.573 7.90e-06 ***
## full_teamDelta Team 6  -11.947      1.915  -6.240 2.12e-09 ***
## full_teamGamma Team 1  -11.802      2.553  -4.623 6.34e-06 ***
## full_teamGamma Team 2   -8.771      1.977  -4.436 1.43e-05 ***
## full_teamGamma Team 3   -7.067      2.553  -2.769 0.006096 **
## full_teamGamma Team 4  -13.417      1.789  -7.502 1.41e-12 ***
## full_teamGamma Team 5  -11.708      1.789  -6.546 3.90e-10 ***
## full_teamGamma Team 6  -13.714      1.712  -8.009 6.01e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.699 on 227 degrees of freedom
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.4736
## F-statistic: 10.78 on 23 and 227 DF,  p-value: < 2.2e-16
```

```
anova(w1_int, w1_full)
```

```
## Analysis of Variance Table
##
## Model 1: productivity ~ duty_shift * shift_team
## Model 2: productivity ~ full_team
##   Res.Df    RSS Df   Sum of Sq F Pr(>F)
## 1    227 3106.4
## 2    227 3106.4  0 -2.2737e-12
```

Timed assessment question

Consider the models w1_shift, w1_team, w1_int and w1_full and their summaries. Choose the most appropriate model to help you answer the following.

Suppose you, as Chief Science Officer, are assigned to Alpha shift Team 1 (you're not in this dataset, though). What was the average productivity of the group of colleagues you are assigned to work with during the first week after shore leave? Round your answer to the nearest whole number, e.g. 17 or 42.

- **Your answer:** 41.
- **Correct answer:** 41.

Well done, 1 point.

Question 7 (2 points)

There were two versions to this question.

Consider the p-value in the final line of the summary for [w1_team] OR [w1_shift] .

Which TWO of the following statements are TRUE based only on this p-value and test? (I.e., not on the coefficients or other parts of the output.)

(Don't choose more than two, but if you're unsure, it is better strategically to only tick the ones you're confident in as you will lose part marks for ticking incorrect options.)

Your answers: This p-value is also exactly what we would expect to get from conducting a one-way ANOVA on shift_team. Only ONE of the other statements is true.

Correct answers:

- This p-value is also exactly what we would expect to get from conducting a one-way ANOVA on shift_team.
- At the 5% p-value threshold, we would reject the hypothesis…

0 point(s).

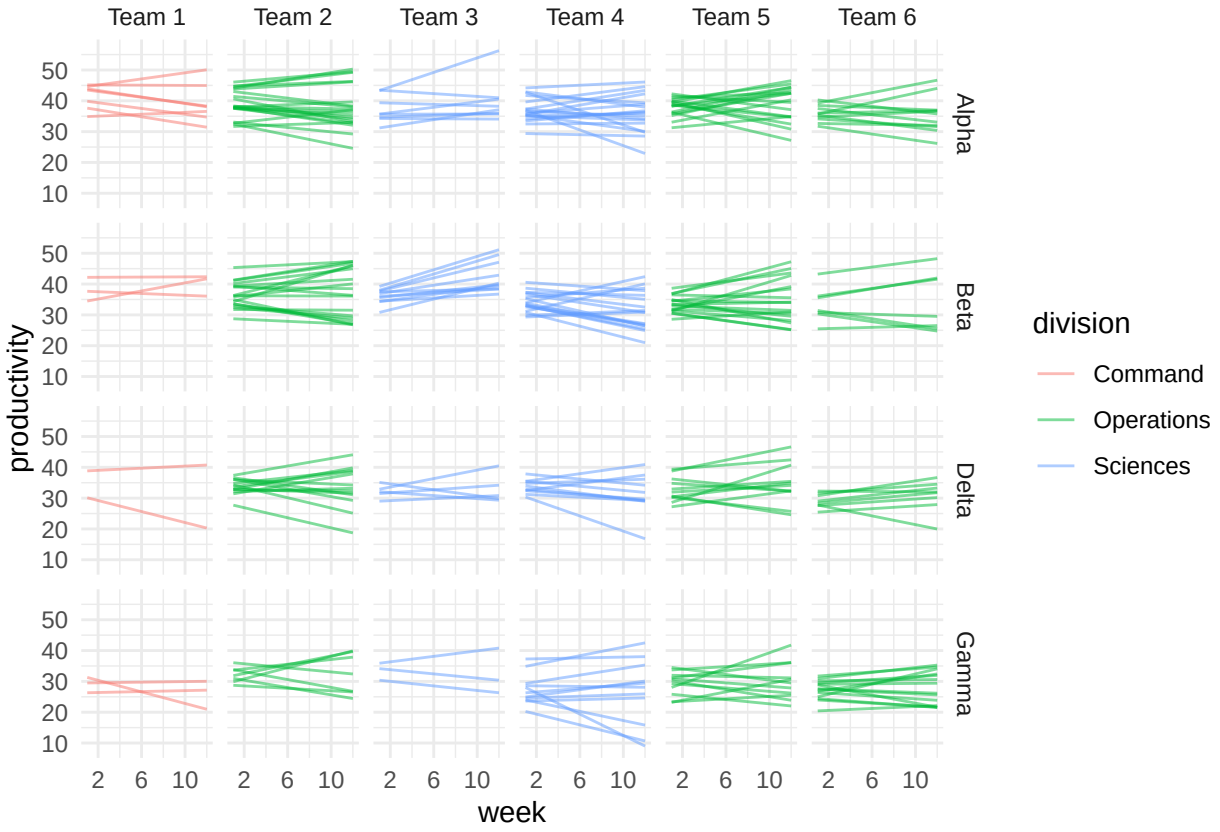
Question 8 (1 point)

Prep from untimed component (task 4, part 1)

Replace the 1s and add whatver other aesthetics are required in the aesthetic statement in the `ggplot()` function to recreate the graph below for your particular ship. Note that each line represents the productivity trend for one crewmember over the 12 week period.

```
# ANSWER
crew_data %>%
  ggplot(aes(y = productivity, x = week, group = crew_id, colour = division)) +
  geom_line(stat="smooth",method = "lm", alpha = 0.5) +
  facet_grid(duty_shift~shift_team) +
  scale_x_continuous(breaks = seq(2,12, by = 4)) +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Timed assessment question

Which ONE of the following is the correct and complete set of aesthetics to satisfy part 1 of task 4? That is, which one of the following sets of aesthetics could be used to help generate the required faceted plot?

- **Your answer:** y = productivity, x = week, group = crew_id, colour = division
- **Correct answer:** `y = productivity, x = week, group = crew_id, colour = division`

Well done, 1 point.

Question 9 (2 points)

Prep from untimed component

After discussing your investigation and the above graph with your Personnel Officer, they suggest you should *not* include rank, position, division, sub-division or gender in your analysis. They also tell you that ship-to-ship, how duty shifts are set up and how teams are allocated differs quite a lot. Some ships have more than the 4 shifts yours does, or have many more teams due to size, etc.

You're interested in presenting your work at the next Federation Science and Innovation Conference and want be able to provide information that might be relevant to the the Chief Science Officers on other ships, too.

Below are several models that you've fit and some tests on them.

```r
# Provided code, changed slightly to better apply the
# optimizer options
control.opts <- lmerControl(optCtrl=list(xtol_abs=1e-8,
                        ftol_abs=1e-8,
                        optimizer = "Nelder_Mead"))

model_1a <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                   (1|name), control = control.opts,
                 data = crew_data)

model_1b <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                   (1 + week|name),  control = control.opts,
                 data = crew_data)

# Study prompt: How do we interpret the p-values here? What is relevant?
lmtest::lrtest(model_1a, model_1b)
```

```
## Likelihood ratio test
##
## Model 1: productivity ~ week + starfleet_gpa + perseverance_score + (1 |
##     name)
## Model 2: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -7016.7
## 2   8 -5421.3  2 3190.8  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Provided code
model_2a <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                   (1 + week|name) + (1|duty_shift:shift_team),
                   control = control.opts,
                 data = crew_data)

model_2b <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                   (1 + week|name) + (1|full_team),
                   control = control.opts,
                 data = crew_data)

# Study prompt: How do we interpret the p-values here? What is relevant?
lmtest::lrtest(model_1b, model_2a)
```

```
## Likelihood ratio test
##
## Model 1: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name)
## Model 2: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name) + (1 | duty_shift:shift_team)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -5421.3
## 2   9 -5345.8  1 150.87  < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmtest::lrtest(model_2a, model_2b)
```

```
## Likelihood ratio test
##
## Model 1: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name) + (1 | duty_shift:shift_team)
## Model 2: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name) + (1 | full_team)
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   9 -5345.8
## 2   9 -5345.8  0     0          1
```

Timed assessment

There were two versions for this question Consider [model_x] and [model_x] and the tests on them. Which ONE of the following is the best conclusion?

- **Your answer:** At the 5% level, we reject the hypothesis that adding a random slope across weeks for each crewmember doesn't help explain the data better.
- **Correct answer:** At the 5% level, we reject the hypothesis that adding a random slope across weeks for each crewmember doesn't help explain the data better.

Well done, 2 points.

Question 10 (1 point)

Prep from untimed component (task 4, part 2)

Determine which model from the above is the most appropriate out of those shown. Make appropriate alterations to `model_3` so that it will be the same as your chosen model with the addition of the term shown below, and uses the appropriate likelihood method to allow you to compare the models.

```
# Provided code
model_3 <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                  (1 + week|name) +
                 (1 + week|full_team),
                 control = control.opts,
                data = crew_data)

lmtest::lrtest(<your chosen model name here>, model_3)
```

```
# ANSWER
model_3 <- lmer(productivity ~ week + starfleet_gpa + perseverance_score +
                  (1 + week|name) +
                 (1 + week|full_team),
                 control = control.opts,
                data = crew_data)

lmtest::lrtest(model_2b, model_3)
```

```
## Likelihood ratio test
##
## Model 1: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name) + (1 | full_team)
## Model 2: productivity ~ week + starfleet_gpa + perseverance_score + (1 +
##     week | name) + (1 + week | full_team)
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   9 -5345.8
## 2  11 -5345.1  2 1.375     0.5028
```

Timed assessment question

Of the models shown in part 2 of task 4, which model or models is/are the most appropriate (or equally most appropriate) to present at the Federation Science and Innovation Conference? Tick all that apply.

- **Your answer:** Model 2a,Model 2b
- **Correct answer:** Model 2a,Model 2b

1 point(s).

Question 11 (3 points)

Prep from untimed component (task 4, part 3)

Run `summary()` and `confint()` on whichever model you think is the most appropriate

```
# ANSWER

# models 2a and 2b are equivalent, one just makes the interaction in the model explicit
# whil the other by uses the variable 'full_team' to represents the interaction

summary <- summary(model_2b)
confint <- confint(model_2b)
```

```
## Computing profile confidence intervals ...
```

```
confint
```

```
##                          2.5 %     97.5 %
## .sig01              2.62813962  3.18558565
## .sig02             -0.23482667  0.03521713
## .sig03              0.46047552  0.55127300
## .sig04              2.48106689  4.62289007
## .sigma              0.97424589  1.02967038
## (Intercept)         8.23812424 16.70078136
## week               -0.05014207  0.07641597
## starfleet_gpa       1.37737401  2.21326863
## perseverance_score  0.79847388  1.53152891
```

Timed assessment question

Which of the following conclusions could you correctly draw from your final most appropriate model? Tick ALL that apply. (Note that you will lose part marks for ticking incorrect options.)

- Your answer: Crewmembers with higher GPAs upon graduating from Starfleet Academy tend to have higher productivity.,On average, crewmembers with higher perseverance scores are more productive.

Statement	Truth value	Contribution to your final score
A crewmember's GPA upon graduating from Starfleet Academy does not appear to be related to their productivity.	FALSE	1
Crewmembers with higher GPAs upon graduating from Starfleet Academy tend to have higher productivity.	TRUE	1
Over the 12 weeks after shore leave, productivity decreases statistically significantly.	FALSE	1
There is no evidence of changing productivity over time in the 12 weeks after shore leave.	TRUE	-1
On average, crewmembers with higher perseverance scores are more productive.	TRUE	1
None of these conclusions are appropriate.	FALSE	1

Your score for this question is 2 points.

Prep from untimed component (task 5)

While on shore leave, some of the astrobiologists had a little competition to see who could spot plants from the greatest number of **different planets or systems** in the hotel gardens. Note: The *number* of plants spotted doesn't actually matter as long as at least one was spotted.

They have asked for your impartial help to find out who the winner is.

You have three datasets:

- `astrobiologists` is a list of all the astrobiology crewmembers
- `competition_data` has the number of plants of each type that each participating astrobiologist recorded.
- `origin_data` contains information from the hotel about the plants in their collection and the the planets they are native to. They have warned you that is may be somewhat incomplete.

Tip: I recommend run `View()` on `competition_data` and `origin_data` to explore them further so you are familiar with their structure and contents. (You can also do this by clicking on their titles in the Environment pane.)

1. Create a new dataset called `complete_comp` using the `competition_data`.
2. Assess whether `complete_comp`, at this current step, is currently tidy. (I.e., is `competition_data` tidy?) If yes, proceed. If no, alter it to be tidy. Specifically, it needs to be in the correct format to be useful for merging the `origin_data` on to it.
3. Continuing to manipulate the `complete_comp` object, merge on the `origin_data` such that any plants **not** present in the data provided by the hotel are **dropped**.

4. Restrict the `complete_comp` so it only contains rows where at least one plant was spotted.
5. Restrict the `complete_comp` to just observations from distinct planets or systems for each crewmember. (See hint code below.)
6. Calculate how many unique planets or systems each astrobiologist spotted at least one plant from.

You DO NOT have to use the exact same code I do to get the associated questions in the timed component correct, as long as it fulfills these instructions, in the correct order. As a hint, here is the structure of my code to complete these tasks.

```
complete_comp <- competition_data %>%
  ____________________ %>%
  ____________________ %>%
  ____________________ %>%
  distinct(crewmember, native_to) %>% # this line will achieve instruction 5
  ____________________ %>% # these last two lines achieve instruction 6,
  ____________________ # but could be done in only one line also
```

```
# ANSWER
complete_comp <- competition_data %>%
  pivot_longer(-crewmember, names_to = "plant", values_to = "count") %>%
  right_join(origin_data, by = "plant") %>%
  filter(count > 0) %>%
  distinct(crewmember, native_to) %>%
  group_by(crewmember) %>%
  summarise(n = n())

complete_comp_2 <- complete_comp %>%
    arrange(desc(n)) %>%
    slice(1) %>%
    mutate(crewmember = str_remove_all(crewmember, "[^A-Za-z -]")) %>%
    mutate(crewmember = trimws(crewmember))
```

Question 12 (2 points)

- **Your answer:** This dataset is not tidy because it is not the case that each variable has one and only one column.
- **Correct answer:** This dataset is not tidy because it is not the case that each variable has one and only one column.

Well done, 2 points.

Question 13 (2 points)

- **Your answer:** right_join
- **Correct answer:** `right_join`. Only a right join would work exactly as described in the instructions.

Well done, 2 points.

Question 14 (1 point)

- **Your answer:** Kawthar al-Sharaf (42126)
- **Correct answer:** Kawthar al-Sharaf

Well done, 1 point.

Question 15 (3 points)

Prep from untimed component (task 6)

Suppose you were trying to run the following code. It throws an error. (Note: DON'T fix the error, that isn't the point of this activity.) Create a reprex (a reproducible example, see week 1) with everything required for your statistician to reproduce this error. The only 'error' in the output should be the one produced by *this* code. (Hint: there is a library you should include, and you'll also need to provide the data. Once you've copied the complete code for the reprex to your clipboard, you can then run `reprex()` and the content for your reprex will then be added to you clipboard, (i.e., with Ctrl+V or Cmd+V you can paste it.))

```
origin_data %>%
  filter(nativeto == "Delta Quadrant")
```

```
# ANSWER

# a full answer needs the library (1 point),
# the data (1 point) and
# incorrect code to produce the correct error (1 point)

library(tidyverse)

origin_data <- data.frame(plant = c("Xupta tree","L'maki","Leola root",
                                    "Waterplum","Vulcan orchid",
                                    "Lunar flower","Garlanic tree",
                                    "Folnar jewel plant",
                                    "Felaran rose","Crystilia","Parthas",
                                    "Borgia plant","Pod plant"),
                          native_to = c("Orellius system","Delta Quadrant",
                                    "Bajor","Mari","Vulcan",
                                    NA,"Elaysian homeworld","Folnar III",
                                    "Delta Quadrant","Telemarius IV",
                                    "Acamar III","M-113",NA))

origin_data %>%
  filter(nativeto == "Delta Quadrant")
```

Mark scheme

- `library(tidyverse)` included: 1
- code to create data included: 1
- correct error reproduced: 1

Your score: 3 point(s).

Question 16 (2 points)

- **Your answer:** We can scrape this part of the site as long as we use a 5 second crawl delay, take only what we need, and credit the original source.
- **Correct answer:** Scraping this part of the site is not allowed without permission from the site.

Incorrect.

Question 17 (2 points)

- **Your answer:** Case-control study
- **Correct answer:** Case-control study,

Well done, 2 points.

Question 18 (3 points)

- **Your answer:** There may be unmeasured confounders that could distort our understanding of the association between platform (console or computer) and rating.,To achieve the goals of this investigation, we should consider fitting a random effect for customer (i.e., use email as an ID) to account for repeated ratings measures.
- **Correct answers:**
  - There may be unmeasured confounders that could distort our understanding of the association between platform (console or computer) and rating.
  - We should not use linear regression to fit this model.
  - We should not use a linear mixed model for this investigation.

Your score: 0 point(s).

Final score: 23/31 = 74.2%