STA130 Rstudio Homework **Solutions**

Problem Set 9

[Xuanqi Wei] ([1009353209]), with Josh Speagle & Scott Schwartz

```
student_num =   1009353209
student_num_last3 =  209
```

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Thursday, March 30.

Question 0: Ethics in Data Science

A number of ethical concerns often arise when conducting research with people. Consider the following two situations.

(a) A data analyst receives permission to post a data set that was scraped from a social media site, provided it is appropriately "de-idenfitied" (i.e. ensure that each entry cannot be linked to a specific individual). The full data set included:

- name,
- screen name,
- email address,
- geographic location,
- IP (internet protocol) address,
- demographic profiles, and
- preferences for relationships.

The researcher believes just removing the name and email address should be enough to de-identify the data. Are they correct? If not, what issues might still be present?

Note: This question is taken from Section 8.12, Problem 8 in "Modern Data Science with R (2nd edition)".

> Yes, they are correct because besides name and email address, there are other personal information that can help us identify the person. But with the name and email address, the identification is definitely easier.

(b) A reporter carried out a clinical trial of chocolate where a small number of overweight subjects who had received medical clearance were randomized to either eat dark chocolate or not to eat dark chocolate. They were followed for a period and their change in weight was recorded from baseline until the end of the study. More than a dozen outcomes were recorded and one proved to be significantly different in the treatment group than the outcome. This study was publicized and received coverage from a number of magazines and television programs. What are some of the potential ethical considerations that could arise in this situation?

Note: This question is adapted from Section 8.12 Problem 5 in "Modern Data Science with R (2nd edition)".

> The potential ethical considerations that could arise in this situation could be informed consent as they might not be informed to be included in the study. Moreover, the problem in misrepresentation of study result as well.

Could there be additional statistical concerns due to the fact that only a small number of subjects were included in the study and a large number of outcomes were tracked and recorded? Why or why not?

> The could be additional statistical concerns. Because only a small number of subjects were included can cause lack of data in the study and a large number of outcomes can cause the poor analysis issue, which both will lead to wrong result.

Introduction to Lumosity

For the remainder of the assignment we will focus specifically on some of the more subtle (or perhaps not so subtle) ways in which various issues can impact our results and/or our interpretation of our results. Note that most of the analysis methods should be very closely related to methods and code you have implemented in previous problem sets, and so the difficulty/length of these questions should be quite a bit easier/shorter than they might appear at first glance if you can take full advantage of your past work.

Lumosity is a brain training app that claims to help improve cognitive skills such as memory, reasoning and focus. A large randomized trial was conducted to evaluate the impact of Lumosity training on cognitive skills. The study and results are presented in Hardy et al. (2015)'s publication *"Enhancing Cognitive Abilities with Comprehensive Training: A Large, Online, Randomized, Active-Controlled Trial"*.

In the study, thousands of participants were recruited from Lumosity's free users (i.e., people who set up free Lumosity accounts but did not pay for full access) and randomly assigned to one of two groups:

- **Treatment Group** (Lumosity Training): Participants in the treatment group completed Lumosity training online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks.
- **Control Group** (Crossword Puzzles): Participants in the control group completed crossword puzzles online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks.

Looking at the main differences between the two groups, the main variable being tested was whether doing Lumosity training exercises on a regular basis were more effective than the "baseline" activity of doing crosswords for roughly the same amount of time and at the same frequency. In other words, it's not trying to measure whether Lumosity training is better than *nothing*, but whether it's better than another simple yet popular activity (crossword puzzles) that many people do on a regular basis. Note that regular crossword puzzle solving has itself been associated with numerous claims of improved cognitive performance and have also been studied in quite some detail, which is likely one reason why it was used as a control activity for this study.

Let's now load in the data.

```
library(tidyverse)

# load in the data
study_dat <-
  read_csv("lumosity_study_data.csv") %>%
  na.omit()
glimpse(study_dat)
```

```
## Rows: 4,976
## Columns: 6
## $ participant_id     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
## $ group              <chr> "crosswords", "lumosity", "crosswords", "crosswords~
## $ age_round          <dbl> 51, 24, 19, 21, 31, 25, 45, 27, 40, 50, 46, 52, 21,~
## $ GI_improve         <dbl> -5.992334, 9.030539, -15.030917, -9.265834, 3.18986~
## $ concentration_post <dbl> 1, 3, 5, 2, 4, 1, 4, 5, 5, 4, 4, 4, 5, 4, 4, 2, 4, ~
```

```
## $ active_days        <dbl> 56, 54, 6, 21, 43, 15, 0, 40, 39, 69, 64, 32, 21, 4~
```

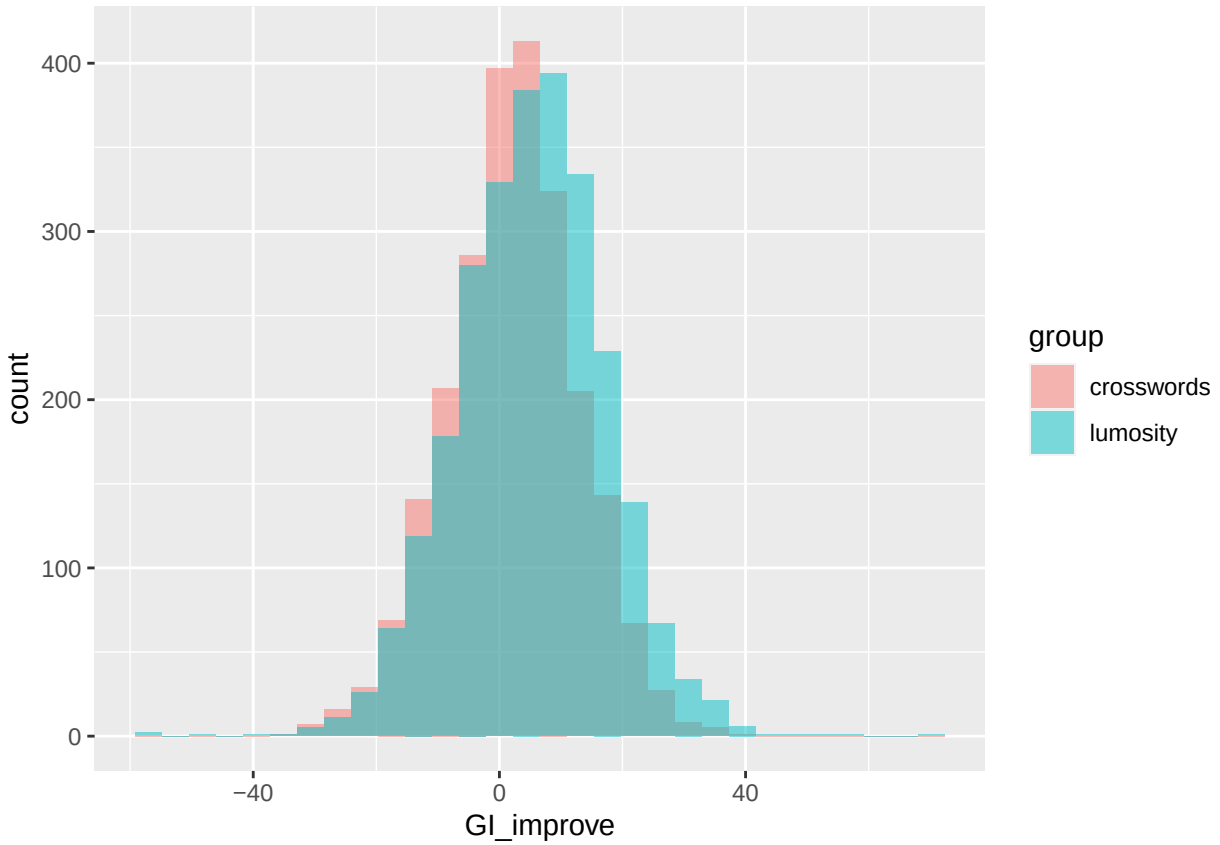
This data set contains information on the 5045 Lumosity users who participated in the study. The main measure of cognitive skills was called the Grand Index (GI) Score, and higher GI values imply better cognitive skills. The cognitive skills of the participants who completed the study were scored before and after the 10-week study period, with the difference recorded as the `GI_improve` variable.

Let's visualize the distribution of the `GI_improve` values between the two groups now.

```
# generate summary table
group_by(study_dat, group) %>%
  summarise(n=n(),
            mean=mean(GI_improve),
            sd=sd(GI_improve),
            median=median(GI_improve),
            IQR=IQR(GI_improve)
            )
```

```
## # A tibble: 2 x 6
##   group          n  mean    sd median   IQR
##   <chr>      <int> <dbl> <dbl>  <dbl> <dbl>
## 1 crosswords  2346  2.14  10.6   2.37  14.0
## 2 lumosity    2630  5.28  12.0   5.56  15.7
```

```
# generate grouped boxplot comparison
study_dat %>%
  ggplot() +
  aes(x=GI_improve, fill=group) +
  geom_histogram(position="identity", alpha=0.5)
```



Several other potentially useful variables from the study are also included including the participants' ages (`age_round`), ability to concentrate ranked from 1-5 (`concentration_post`), and the number of active days participating in study activities (`active_days`).

Question 1: Lumosity, Crosswords, and Cognitive Performance

Let's now investigate the central question of the study: does the impact of Lumosity training differ from that provided by solving crossword puzzles?

(OPTIONAL) (a) Perform a **two-sample permutation hypothesis test** using $m = 1000$ trials to compare whether the **mean** Grand Index score improvements after online training (`GI_improve`) between the `lumosity` ($n = 2667$) and `crosswords` (2378) groups are the *same* (null hypothesis) or *different* (alternative hypothesis) using an $\alpha = 0.05$ rejection rule. Plot your resulting sampling distribution and also print out your computed p-value and the result of your hypothesis test after applying your rejection rule.

Hint: Your code from HW4 may be helpful here!

```
set.seed(student_num_last3)  # required to ensure reproducibility

# code your answer here
```

(b) Based on your 2-sample permutation test above, **you should have found that the estimated p-value is 0** (or extremely close to it) and therefore rejected the null hypothesis in favour of the alternative. This naively implies that there is very strong evidence against the null hypothesis that the mean Grand Index Score improvement is the same for those undergoing Lumosity training and those completing online crossword

puzzles.

Do you believe this result? If so, why? If not, why not?

> Yes, I believe this result as the p-value we found is 0 which we have strong evidence to against H0, that the mean Grand Index Score improvement is the same for those undergoing Lumosity training and those completing online crossword puzzles.

(OPTIONAL) (c) Use **bootstrapping** with $m = 1000$ trials to estimate the 95% **confidence interval** for the true population mean of `GI_improve` for the populations associated with each of the groups (`lumosity` and `crosswords`).

Hint: Your code from HW5 may be helpful here!

```
set.seed(student_num_last3 + 1)  # required to ensure reproducibility

# code your answer here
```

Comment on whether your estimated 95% CIs support or weaken the results of your two-sample hypothesis test.

> *REPLACE THIS TEXT WITH YOUR ANSWER*

Question 2: Study Design

Outside of any potential issues you may have discussed above, there might also be issues with the inherent study design used by Hardy et al. (2015).

(a) Take a look at Figure 1 from the Hardy et al. (2015) paper, which summarizes their study design. In 1-3 sentences, describe what this figure tells us about how the study was performed and what data was eventually used for the comparison we made earlier.

> 1. Initially, a cohort of 11,470 individuals was recruited, from which 1,551 participants were excluded on the basis of various factors, including age.
> 2. Subsequently, the remaining 9,919 individuals were divided into two groups using a control variable method. During the course of the experiment, attrition occurred, with 2,384 participants lost to follow-up in the treatment group, 2,490 participants lost to follow-up in the control group, and 330 participants excluded due to active cognitive training.
> 3. As a result, the final sample size comprised 2,667 individuals in the treatment group and 2,048 individuals in the control group.

(b) Is the data we analyzed above (and got a p-value of 0 for) the full data of a randomized controlled trial? Or a subset of data that is based on individual behavior (and hence in some sense "observational")?

- The data we analyzed above (and got a p-value of 0 for) the full data of a randomized didn't controll trial, it's just a subset of data that is based on individual behavior (and hence in some sense 'observational').

(c) With these pieces of information in mind, can we conclusively conclude based on these results that Lumosity training leads to more improvement in cognitive skills compared with completing crossword puzzles online? Explain your answer in 1-3 sentences.

> No, with these pieces of informatino in mind, we can't conclusively conclude based on these results. s our analysis is based on a subset of the dataset, there is potential for confounding to exist, as the results may be influenced by self-selection bias, wherein individuals who are more predisposed to participate in the study may be overrepresented in the sample.

Question 3: Two Groups, Multiple Variables

There are multiple additional variables provided as part of the dataset. These may serve as **confounding variables** that also correlate with `GI_improve`. If the general distribution of these variables differs between the two groups, that might explain some (or potentially *all*) of the observed effect.

(a) Let's first consider the participants' ages, as recorded by `age_round`. We might expect that a participant's age correlates with their overall ability to improve their cognitive skills (e.g., older participants may benefit more than younger participants). Do you think user ages would be different between the Lumosity group and crossword groups? Why or why not?

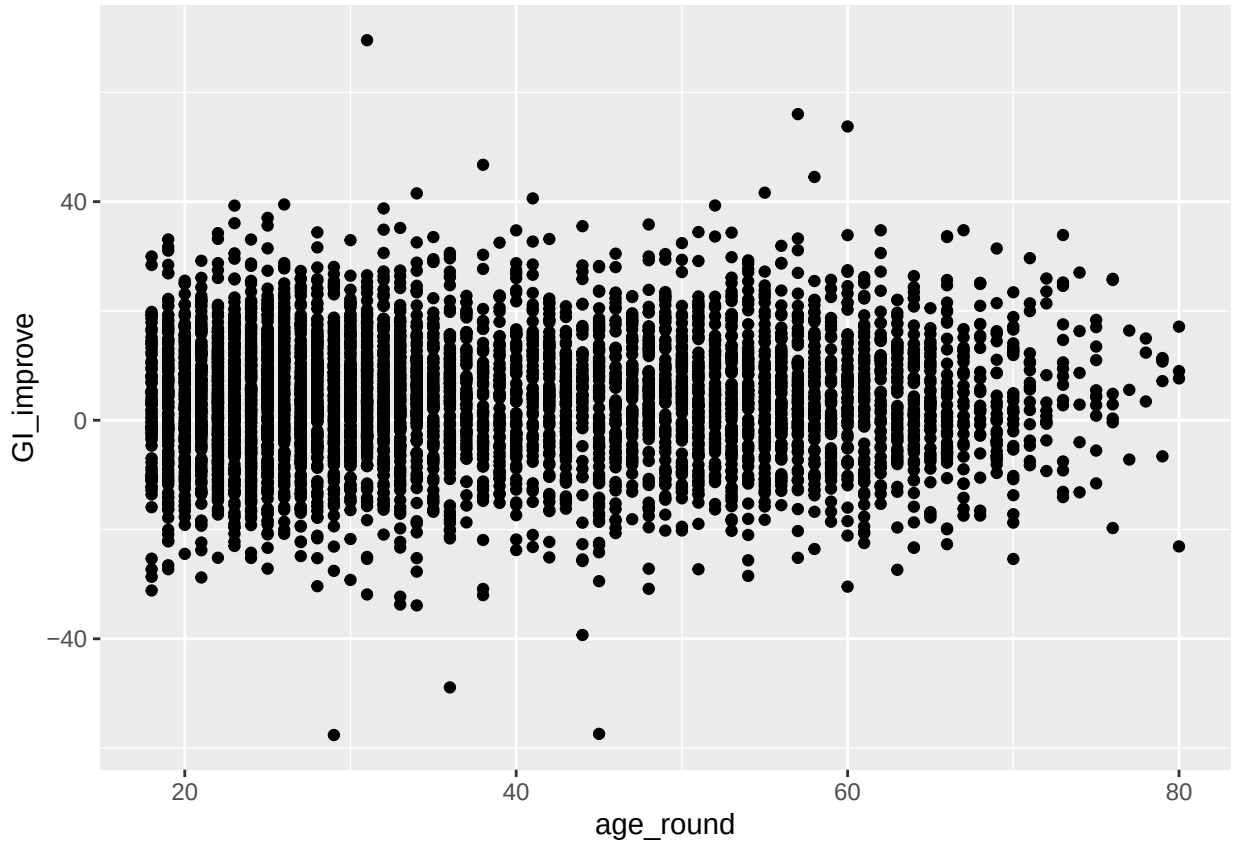
> No, I don't think user ages would be different between the lumosity group and crossword groups because they are randomly chosen.

(b) Produce appropriate data summaries/visualizations to see if:

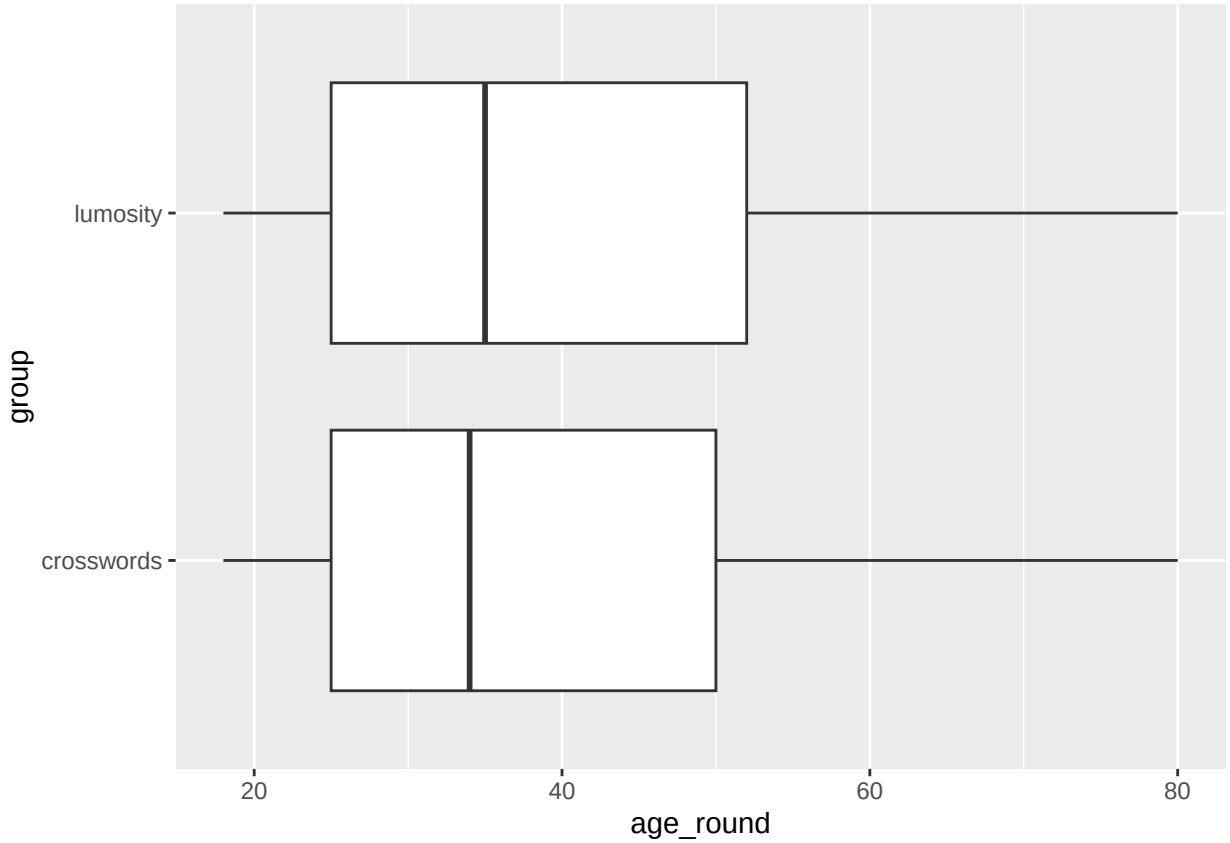
1. the ages (`age_round`) of the users correlates with their Grand Index score differences (`GI_improve`) and
2. whether the ages differ for the Lumosity and crossword groups (`group`).

Hint: Your code from HW2 may be helpful here!

```
study_dat %>% ggplot(aes(x=age_round, y=GI_improve)) + geom_point()
```




```
study_dat %>% ggplot(aes(x=age_round, y=group)) +
geom_boxplot()
```



Interpret your summary and comment on how this compares to your prediction about how ages would compare.

> Comparing the graph illustrations above, there are slightly difference between the ages in two groups, which are basically the same as my prediction.

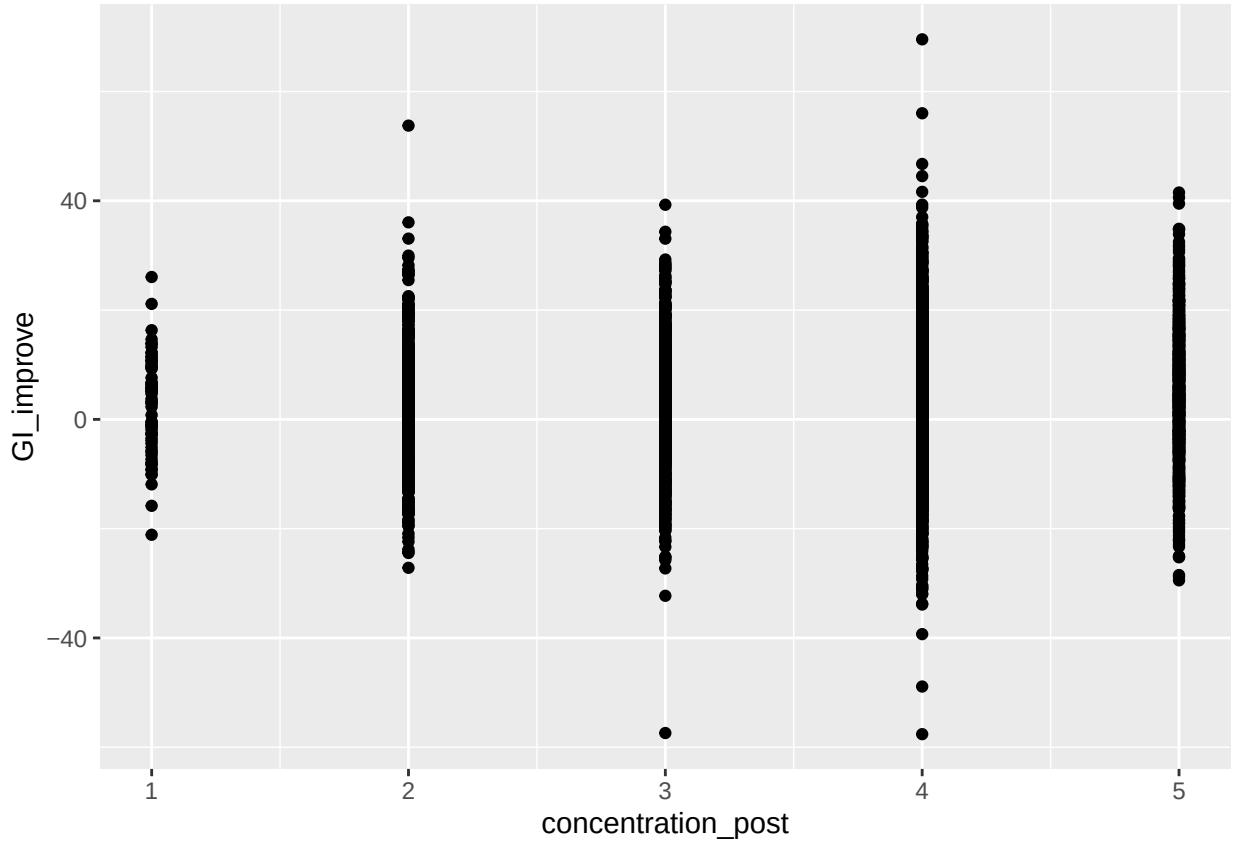
(c) What about the other two variables, `concentration_post` and `active_days`? What effects might these have on `GI_improve`? And would we expect to potentially see differences in their distributions across the two samples?

> The results suggest that concentration_post may be the only significant difference. However, further testing is necessary to confirm the findings and account for potential confounding factors.

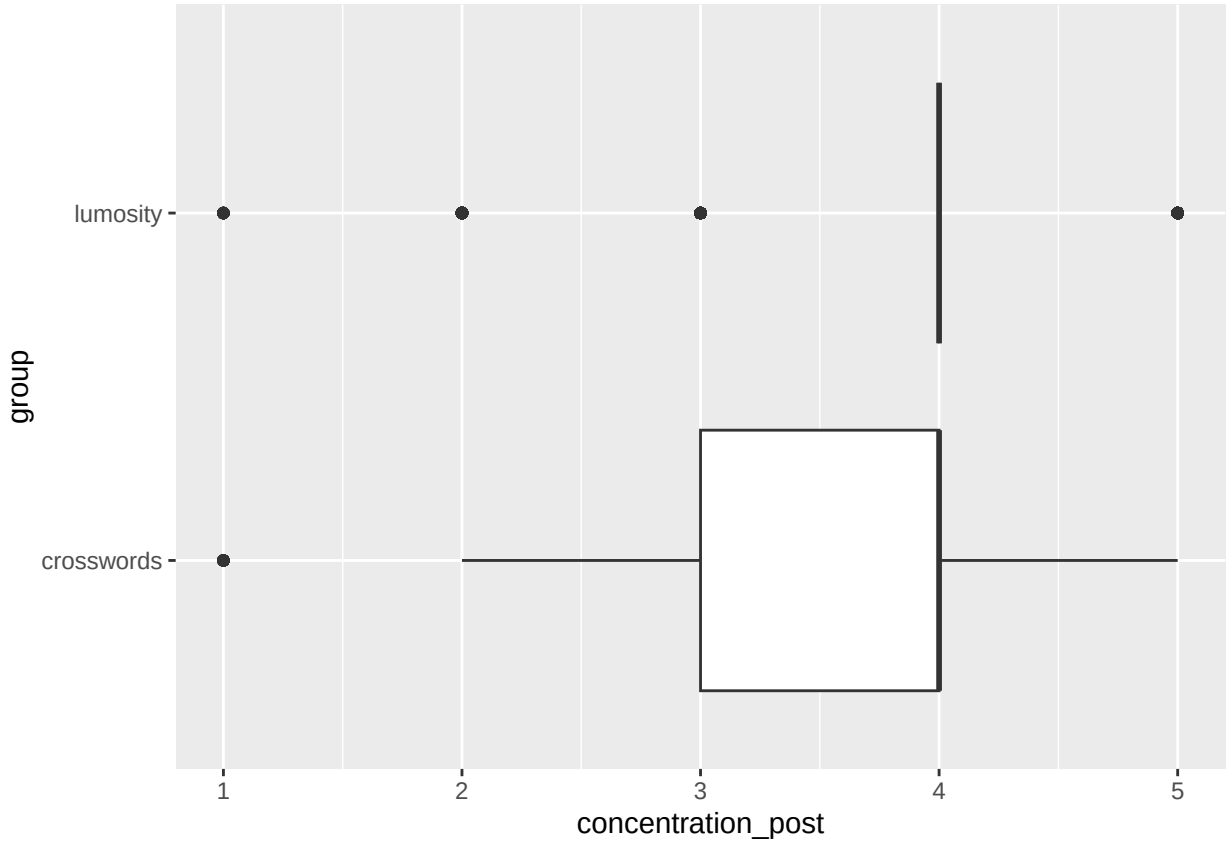
(d) Produce appropriate data summaries/visualizations to see if either of these variables correlates with `GI_improve` and/or if their distributions appear to differ between the two groups.

Hint: Your code from HW2 may be helpful here!

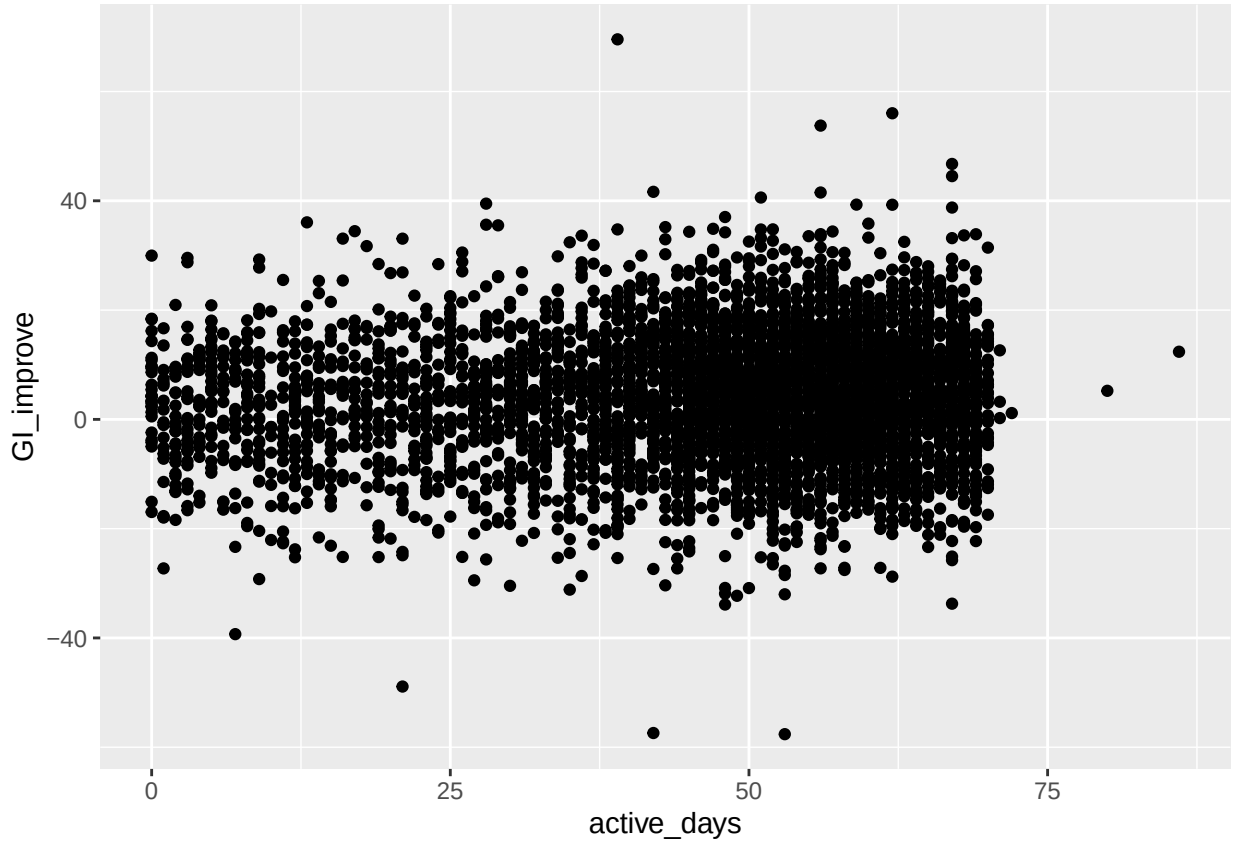
```
study_dat %>% ggplot(aes(x=concentration_post, y=GI_improve)) + geom_point()
```



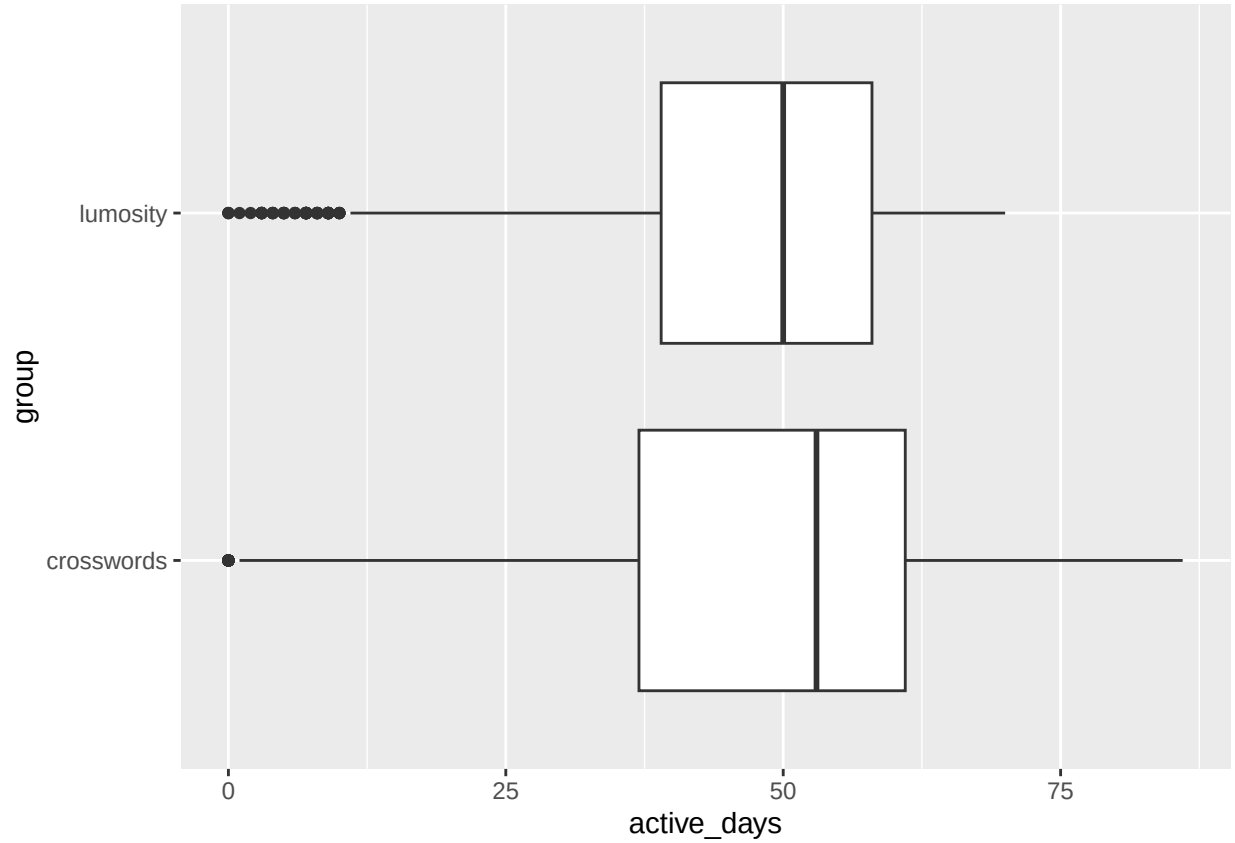

```
study_dat %>% ggplot(aes(x=concentration_post, y=group)) + geom_boxplot()
```




```
study_dat %>% ggplot(aes(x=active_days, y=GI_improve)) + geom_point()
```




```
study_dat %>% ggplot(aes(x=active_days, y=group)) + geom_boxplot()
```



Question 4: Predictions with Multiple Variables

While data visualizations can be useful in diagnosing any obvious issues, a more rigorous way to explore them involves fitting a model that can account for potential effects explicitly.

(a) Let's first explore a simple **linear regression** model. Using the `lm()` function, fit a linear regression model that uses all of the variables available (`group`, `age_round`, `concentration_post`, and `active_days`) **excluding interaction terms**. The print out a summary of the model using `summary()`.

Hint: Your code from HW7 may be helpful here!

```
mod <- lm(GI_improve ~ group+age_round + concentration_post + active_days, data = study_dat)
summary(mod)
```

```
##
## Call:
## lm(formula = GI_improve ~ group + age_round + concentration_post +
##     active_days, data = study_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.107  -7.357   0.263   7.407  64.562
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.747666   0.846062  -0.884 0.376901
```

```
## grouplumosity       3.074526   0.324665   9.470  < 2e-16 ***
## age_round           0.004212   0.011082   0.380 0.703868
## concentration_post  0.270853   0.186865   1.449 0.147272
## active_days         0.037082   0.009811   3.780 0.000159 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.34 on 4971 degrees of freedom
## Multiple R-squared:  0.02275,    Adjusted R-squared:  0.02196
## F-statistic: 28.93 on 4 and 4971 DF,  p-value: < 2.2e-16
```

What is particular coefficient that we can use to test for potential differences in the mean `GI_improve` values between the two samples? Based on the reported test statistic and p-value for this coefficient (from a t-test), should we reject the null hypothesis under the same α significance level as above?

> *Yes, we should reject the null hypothesis under the same α significance level. If the p-value is less than 0.05, it can be inferred that there is significant evidence to support the presence of an association between the variable and GI_improve.*

Is there anything potentially concerning about the above results? If so, what?

> *Yes, the thing potentially concerning about the above results are the confounding variables.*

(b) Repeat the above analysis, but now using a model that **includes all possible interaction terms** (i.e. with the syntax `x1 * x2 * x3 * x4`).

Hint: Your code from HW7 may be helpful here!

```
mod <- lm(GI_improve ~ group * age_round * concentration_post * active_days, data = study_dat)
summary(mod)
```

```
##
## Call:
## lm(formula = GI_improve ~ group * age_round * concentration_post *
##     active_days, data = study_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.364  -7.295   0.278   7.386  64.732
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       4.1345497  6.4665504
## grouplumosity                                     2.2868174  9.7272408
## age_round                                         0.0472572  0.1950325
## concentration_post                               -1.4855465  1.7801465
## active_days                                      -0.0173414  0.1354307
## grouplumosity:age_round                          -0.2183938  0.2845336
## grouplumosity:concentration_post                  0.2832134  2.6216257
## age_round:concentration_post                      0.0064090  0.0524075
## grouplumosity:active_days                        -0.0438772  0.2083241
## age_round:active_days                            -0.0021985  0.0037957
## concentration_post:active_days                    0.0217655  0.0367813
## grouplumosity:age_round:concentration_post        0.0412319  0.0755505
## grouplumosity:age_round:active_days               0.0059654  0.0057009
## grouplumosity:concentration_post:active_days      0.0132979  0.0552459
## age_round:concentration_post:active_days          0.0002681  0.0010130
```

```
## grouplumosity:age_round:concentration_post:active_days -0.0013091  0.0014972
##                                                        t value Pr(>|t|)
## (Intercept)                                              0.639    0.523
## grouplumosity                                            0.235    0.814
## age_round                                                0.242    0.809
## concentration_post                                      -0.835    0.404
## active_days                                             -0.128    0.898
## grouplumosity:age_round                                 -0.768    0.443
## grouplumosity:concentration_post                         0.108    0.914
## age_round:concentration_post                             0.122    0.903
## grouplumosity:active_days                               -0.211    0.833
## age_round:active_days                                   -0.579    0.562
## concentration_post:active_days                           0.592    0.554
## grouplumosity:age_round:concentration_post               0.546    0.585
## grouplumosity:age_round:active_days                      1.046    0.295
## grouplumosity:concentration_post:active_days             0.241    0.810
## age_round:concentration_post:active_days                 0.265    0.791
## grouplumosity:age_round:concentration_post:active_days  -0.874    0.382
##
## Residual standard error: 11.34 on 4960 degrees of freedom
## Multiple R-squared:  0.02598,    Adjusted R-squared:  0.02303
## F-statistic:  8.82 on 15 and 4960 DF,  p-value: < 2.2e-16
```

How do these results compare to your previous ones? Is this surprising? Does it suggest anything regarding the potential impact of confounding variables? If so, why? If not, why not?

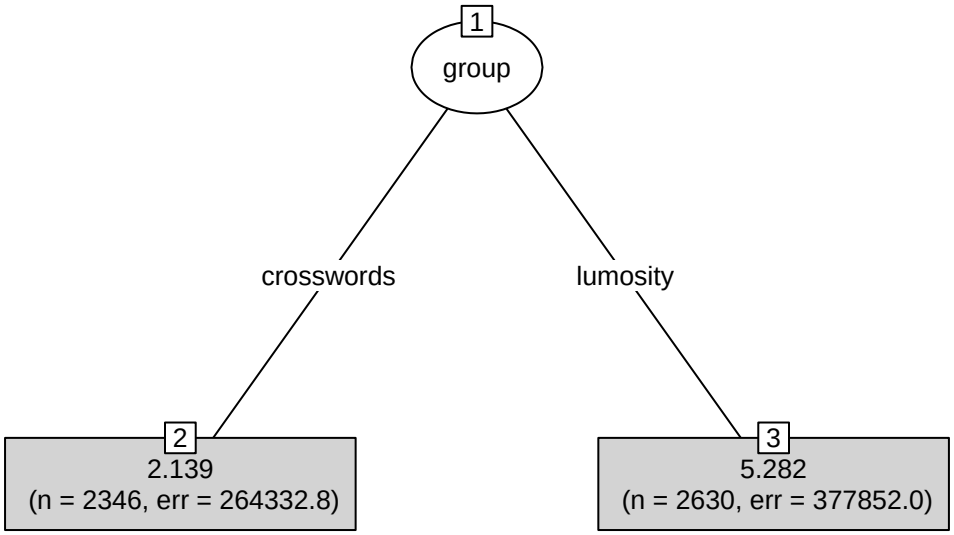
> *Yes, it is surprising. It suggests somethings regarding the potential fonfounding variables as the results have gotten influenced.*

(c) Repeat the above analysis, but now using a **decision tree**. After loading in the `rpart` and `partykit` packages, train a decision tree *using the entire dataset* (i.e. without partitioning it into separate training/validation/testing sets). Then, visualize the resulting tree along with the feature importances.

Note: Using a decision tree this way for a regression problem just involves switching out the loss function to be the mean-squared-error (MSE) and the predictions to be the sample mean of objects within each node, rather than using the Gini impurity as the loss function and having the predictions be the fraction of objects in a given class within each node. This should happen automatically if you use the syntax `rpart(GI_improve ~ x1 + x2)` *based on the data type of* `GI_improve`*.*

Hint: Your code from HW8 may be helpful here!

```r
library(rpart)
library(partykit)

set.seed(student_num_last3 + 2)  # required to ensure reproducibility

tree1 <- rpart(GI_improve ~ group + age_round + concentration_post + active_days, data = study_dat)
plot(as.party(tree1), gp = gpar(cex = 0.8), type = "simple")
```




```
tree1$variable.importance
```

```
##              group        active_days concentration_post          age_round
##        12252.70531          866.98597          783.42106           36.55965
```

How do these results compare to your previous ones? Is this surprising? How might this relate to some of the potential issues with decision trees, such as how splits are calculated, potential splitting rules, and/or issues such as overfitting?

> Yes, this is surprising as well. The observed discrepancy is noteworthy, and potential explanations for the disparity may include differences in splitting rules or overfitting, both of which are recognized issues associated with decision trees.

(d) Repeat the above analysis, but now using a **random forest**. After loading in the `randomForest` package, train a random forest *using the entire dataset* (i.e. without partitioning it into separate training/validation/testing sets). Then, plot a visualization showing the predicted feature importances.

Note: As with the decision tree, the random forest should automatically adjust the loss function and prediction method based on the data type of the input variable (`GI_improve`) even though the input syntax of `randomForest(GI_improve ~ x1 + x2)` is otherwise identical.

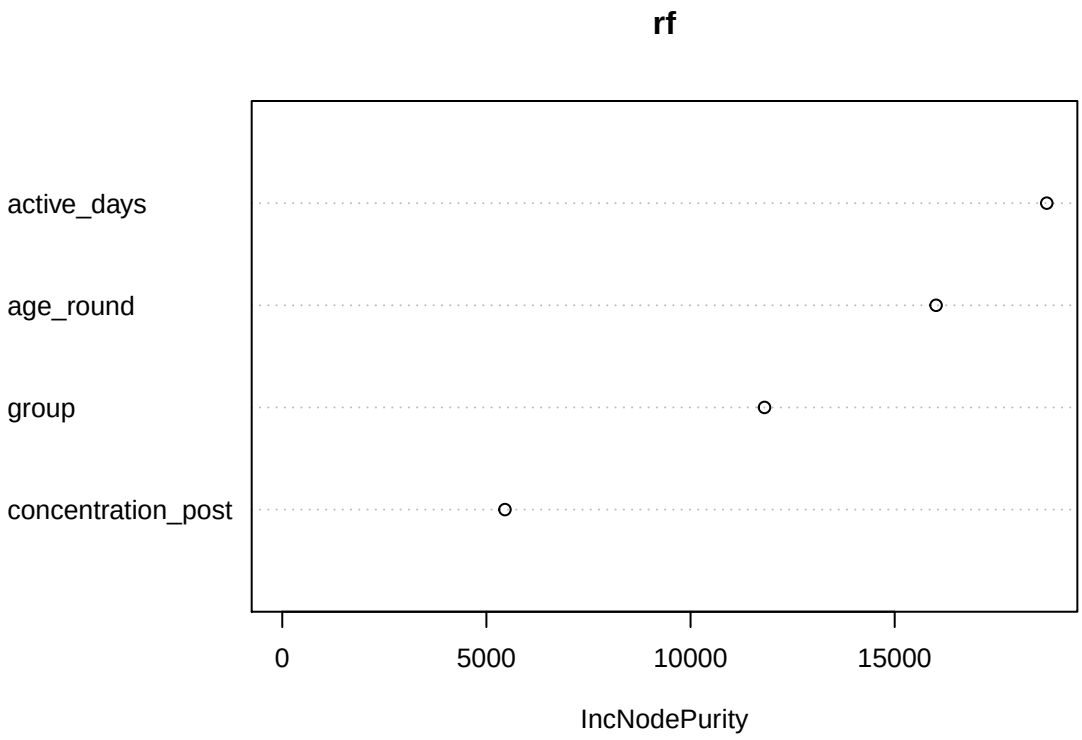
Hint: Your code from HW8 may be helpful here!

```
library(randomForest)

set.seed(student_num_last3 + 3)  # required to ensure reproducibility

rf <- randomForest(GI_improve ~ group + age_round + concentration_post + active_days, data = study_dat)
```

```r
as.vector(rf$importance) / sum(rf$importance)
```

```
## [1] 0.2271259 0.3079279 0.1048834 0.3600627
```

```r
varImpPlot(rf, type = 2, cex = 0.8)
```



How do these results compare to your previous ones? Is this surprising? How might this relate to some of the ways in which random forests potentially address some of the shortcomings of decision trees?

> Yes, these results are suprising compared to the previous one. Specifically, complex random forests are at risk of overfitting, which occurs when the model learns the noise and specific characteristics of the training data, rather than the underlying patterns that are generalizable to new data. Therefore, it is important to carefully balance the complexity of ranfom forest with their ability to generalize to new data.

(e) Based on the above results, what is your overall opinion on whether the impact of Lumosity training on cognitive skills differs from the impact of solving crossword puzzles?

> The overall opinion on whether the impact of lumosity training on cognitive skills differs from the impact of solving corssword puzzles should be related beased on the above results. However, the presence of confounding variables can make it difficult to establish a definitive causal relationship between variables, and caution should be exercised when interpreting statistical findings in such situations.