Week 11 Activity

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
group_data <- read_csv("group_data.csv")
glimpse(group_data)
```

```
## Rows: 329
## Columns: 3
## $ x <dbl> 4.913420, 6.322524, 5.639702, 6.174787, 5.116290, 2.069154, 5.677...
## $ y <dbl> 6.535962, 4.297844, 7.307389, 5.112643, 7.043966, 8.639776, 4.899...
## $ z <chr> "grp02", "grp01", "grp02", "grp01", "grp02", "grp03", "grp02", "g...
```

Part 1

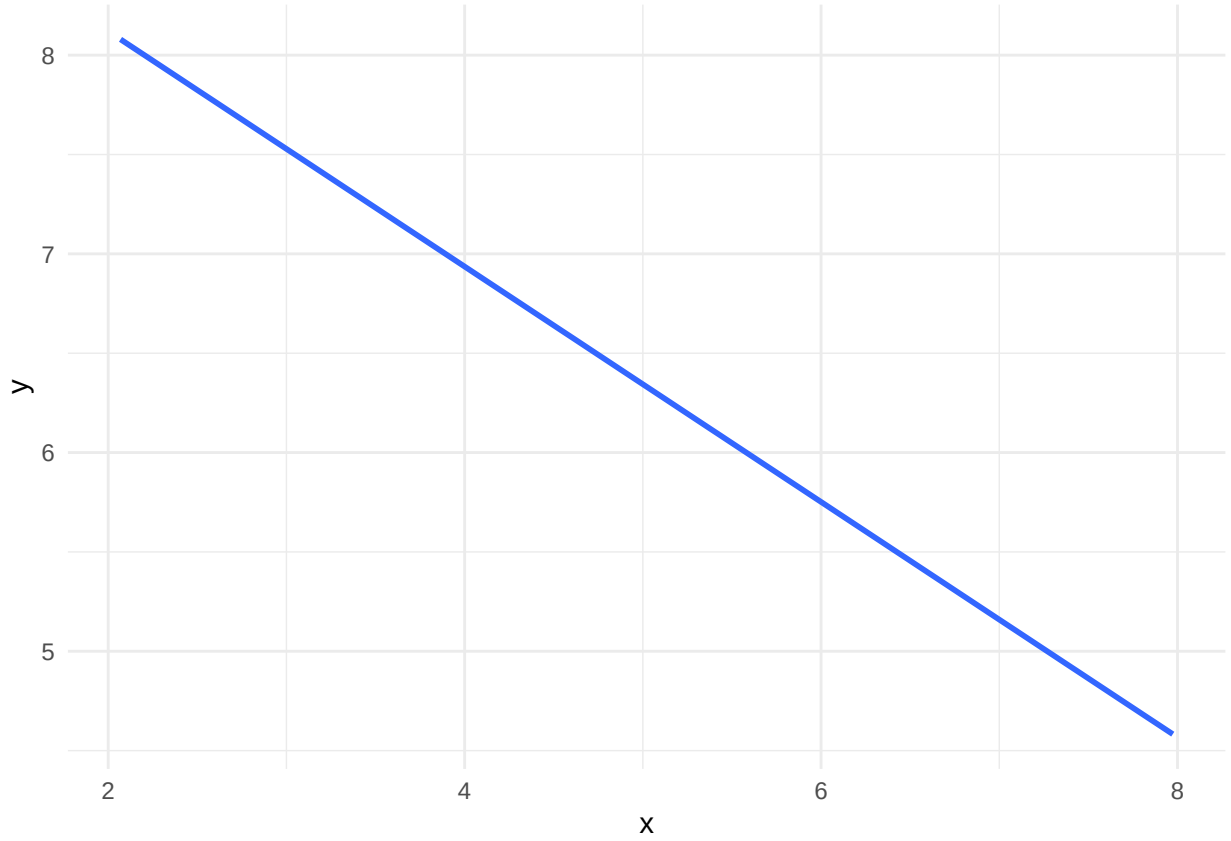
Question 1

What is the best geometry to add to the code below to create an appropriate visualization for the relationship between the variables called 'x' and 'y' in the data set `group_data`?

A. geom_bar()
B. geom_point()
C. geom_line()
D. geom_dot()
E. geom_histogram()

Add it to the code below.

```
group_data %>%
  ggplot(aes(x, y)) +
  # put the correct geometry (i.e. geom_something()) here +
  geom_smooth(method = "lm", se = FALSE, formula = "y~x") +
  theme_minimal()
```



Question 2

Fit the linear model that would give you the equation for the line shown in the plot above and use it to answer the following question. Note: we're not doing a training/test set split here, just use `group_data` as the data.

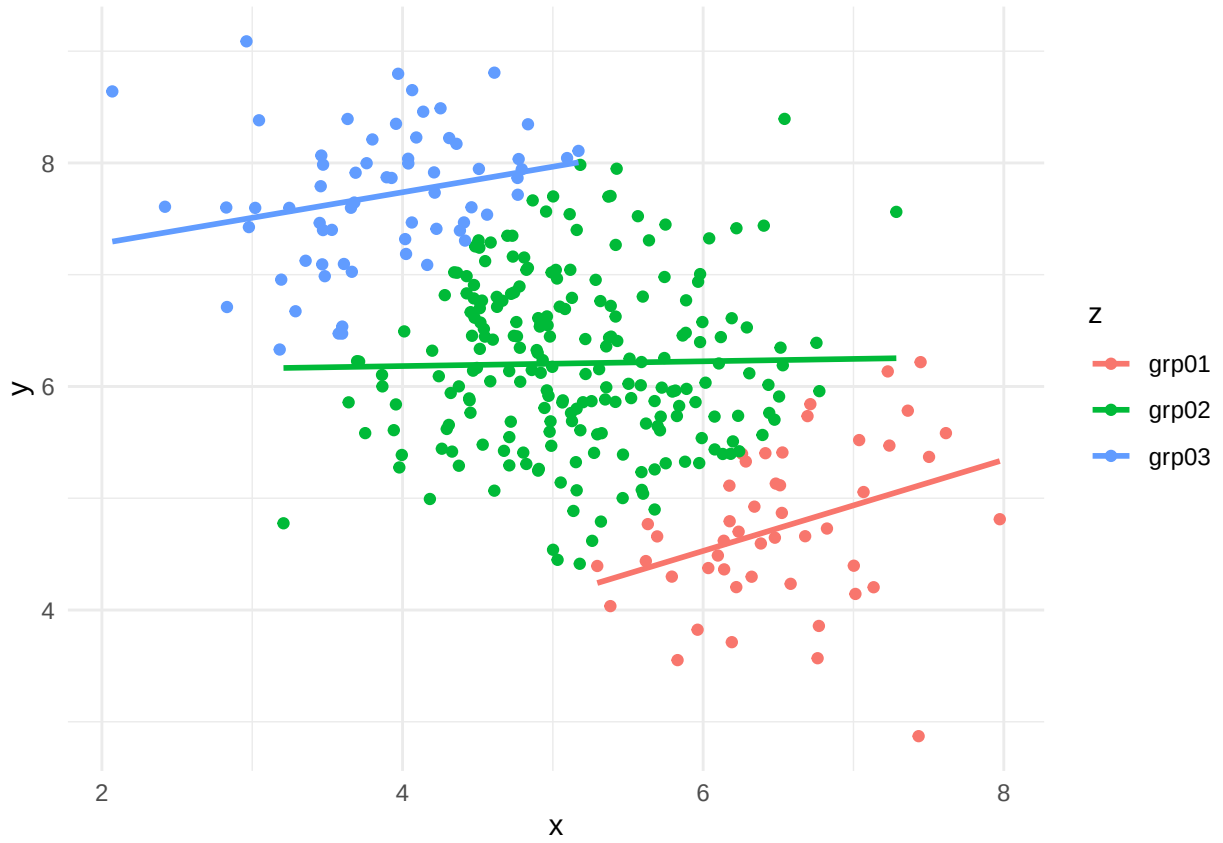
```
# your code here
```

Based on the linear model you fit above, which ONE of the following conclusions is appropriate?

A) For each one-unit increase in `x` we expect `y` to decrease by 0.6 units, on average.
B) For each one-unit increase in `x` we expect `y` to increase by 0.6 units, on average.
C) For each one-unit increase in `x` we expect `y` to increase by 9.3 units, on average.
D) For each one-unit increase in `y` we expect `x` to decrease by 9.3 units, on average.
E) For each one-unit increase in `y` we expect `x` to decrease by 0.6 units, on average.

Question 3

```
group_data %>%
  ggplot(aes(x, y, group = z, colour = z)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = "y~x") +
  theme_minimal()
```



Based on this plot, which of the following statements is FALSE?

A) Within each group, it appears that there is a positive relationship between `x` and `y`.
B) It seems plausible that the slopes for each group may be different.
C) It seems plausible that the intercepts for the lines for each group should be different.
D) As some of the observations in group 1 are very close to the line for group 2, we are likely to have concerns about the RMSE if we were to fit models to training and testing sets and compare.

Question 4

Which of the following chunks of code would fit a linear model that would give you the equations to the three lines above?

A)

```
summary(lm(y ~ x + z, group_data))
```

```
##
## Call:
## lm(formula = y ~ x + z, data = group_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96363 -0.50690 -0.02304  0.51236  2.04550
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.07880    0.39846  10.236   <2e-16 ***
## x            0.10173    0.05895   1.726   0.0854 .
## zgrp02       1.60432    0.14052  11.417   <2e-16 ***
## zgrp03       3.23235    0.20902  15.464   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7192 on 325 degrees of freedom
## Multiple R-squared:  0.5956, Adjusted R-squared:  0.5919
## F-statistic: 159.6 on 3 and 325 DF,  p-value: < 2.2e-16
```

B)

```
summary(lm(y ~ x * z, group_data))
```

```
##
## Call:
## lm(formula = y ~ x * z, data = group_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24125 -0.51551 -0.02164  0.47281  2.15742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0838     1.1284   1.847 0.065707 .
## x             0.4075     0.1722   2.366 0.018561 *
## zgrp02        4.0134     1.1848   3.387 0.000793 ***
## zgrp03        4.7400     1.2558   3.774 0.000191 ***
## x:zgrp02     -0.3861     0.1857  -2.079 0.038370 *
## x:zgrp03     -0.1790     0.2229  -0.803 0.422422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7156 on 323 degrees of freedom
## Multiple R-squared:  0.6022, Adjusted R-squared:  0.596
## F-statistic: 97.78 on 5 and 323 DF,  p-value: < 2.2e-16
```

C)

```
group_data %>%
  group_by(x) %>%
  lm(y ~ x, .) %>%
  summary()
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62923 -0.68463 -0.00669  0.63328  2.96329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.30517    0.26334   35.34   <2e-16 ***
## x           -0.59226    0.05068  -11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.947 on 327 degrees of freedom
## Multiple R-squared:  0.2946, Adjusted R-squared:  0.2924
## F-statistic: 136.6 on 1 and 327 DF,  p-value: < 2.2e-16
```

D)

```
summary(lm(y ~ z, group_data))
```

```
##
## Call:
## lm(formula = y ~ z, data = group_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87125 -0.50834 -0.05226  0.50766  2.18708
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7427     0.1041   45.55   <2e-16 ***
## zgrp02        1.4644     0.1151   12.72   <2e-16 ***
## zgrp03        2.9597     0.1373   21.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7214 on 326 degrees of freedom
## Multiple R-squared:  0.5919, Adjusted R-squared:  0.5894
## F-statistic: 236.4 on 2 and 326 DF,  p-value: < 2.2e-16
```

Question 5

Would you come to generally the same conclusions about the relationship between `x` and `y` from the models in questions 2 and 4? (The correct model from Q4, that is.)

A) Yes, in both there appears to be a positive linear association between `x` and `y`.
B) Yes, in both there appears to be a negative linear association between `x` and `y`.
C) No, in the model from question 2 there appears to be a positive linear association but in question 4' s model there appears to be negative linear association.
D) No, in the model from question 2 there appears to be a negative linear association but in question 4' s model there appears to be positive linear association.
E) Impossible to say.

Question 6

Based on the plots and analyses you've seen in the previous questions, which ONE of the following statements seems most appropriate?

A) We would be concerned about algorithmic bias because the the groups appear to be proxies.
B) There is evidence of disclosure risk due to the strength of association between the variables.
C) There is evidence of over-fitting.
D) The variable `z` appears to confound the relationship between the variables `x` and `y`.

Part 2

Run the following chunk to load some packages for web scraping.

```
library(polite)
```

```
## Warning: package 'polite' was built under R version 4.0.4
```

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.3
```

We're interested in scraping some data from the University of Toronto website. Specifically information on social media accounts.

```
# run this chunk to open the website
browseURL("https://www.utoronto.ca/social-media-directory/all")
```

In the web scraping video for this week, we talked about checking the robots.txt. We can use the `polite` package to check, from within R, whether it seems like this part of the site is allowed to be scraped, based on the robots.txt.

```
session <- bow("https://www.utoronto.ca/social-media-directory/all",
               user_agent = "class activity for STA130, sta130@utoronto.ca")

session
```

Note 1: This chunk is currently set to `eval=FALSE` because unfortunately the output from the package has a character RStudio Cloud's LaTeX can't handle when converting to pdf. You can run this code in the notebook to look at it, no problem.

Note 2: I am setting the `user_agent` to have our contact details in case the web master wanted to get in touch to tell us we were causing issues.

Question 7

Using the output above or by going directly to the robots.txt, what aspect of ethical practice has U of T explicitly asked of us?

A) Not to scrape any of the website at all.
B) Limit the rate at which we scrape the page to 10 seconds per call.
C) Not to crawl any pages that have more than a 10 second loading delay.
D) No specific aspects explicitly asked of us.

Question 8

In addition to checking the robots.txt, what else do we need to consider as ethical scrapers?

A) Check the website's Terms of Use/Terms and Conditions.
B) Check if there is an API available instead.
C) Credit our source.
D) Only take what we need.
E) All of the above.

Below is the code I used to get the data. THIS CODE IS NOT BEING ASSESSED IN STA130. You aren't responsible for understanding it, but I thought some of you might be interested. With the crawl delay it takes a while to run, so I have just given you the data directly in a csv file.

```
pages <- map(1:21, ~scrape(session, query = list(page=.x)) )

social_data <- map_dfr(pages, ~html_node(.x, css = ".view-content") %>%
  html_text() %>%
  str_split(pattern = "\\n") %>%
  unlist() %>%
  as_tibble() %>%
  mutate(value = str_trim(value)) %>%
  filter(value != "") %>%
  mutate(type = if_else(grepl("http", value), "link", "group")) %>%
  mutate(group_name = if_else(type == "group", value, NULL)) %>%
  fill(group_name) %>%
  filter(type == "link") %>%
  select(group_name, value) %>%
  rename(link = "value")) %>%
  mutate(platform = str_remove(link, "https\\:\\/\\/")) %>%
  mutate(platform = str_remove(platform, "http\\:\\/\\/")) %>%
  mutate(platform = str_remove(platform, "www\\.")) %>%
  mutate(platform = str_split_fixed(platform, "\\.", 2)[,1])

write_csv(social_data, "scraped_data.csv")
```

Question 9

What is the most common social media platform used by U of T schools/departments/groups etc.?

A) Facebook
B) Instagram
C) Twitter
D) LinkedIn
E) YouTube

```
social_data <- read_csv("scraped_data.csv")

# your code here
```

Part 3

Question 10

Suppose the Profs post a list of final STA130 grades with student names removed, but include each student's college, gender and tutorial group in the data. Which ONE of the following BEST describes the main issues with this situation? (Also, we would never do this!)

A) This poses issues with informed consent if students were not told their grades would be posted before enrolling in STA130.
B) The poses issues with algorithmic bias as people might use gender to to predict grades for future students.
C) This poses disclosure risk as other members of a tutorial group might be able to identify someone in their tutorial group based on gender and college.
D) This poses an issue for human ethics research as there is no mention of approval from the research ethics board.

Question 11

Suppose when enrolling in STA130 you were randomly assigned to either a mandatory synchronous class or an asynchronous pre-recorded class and then final grades were compared between the two groups. Which ONE of the following is TRUE?

A) Any differences between the groups must be due to random chance because participants were randomly assigned.
B) Random assignment hopefully means the two groups are comparable across potential confounding variables, like high school preparation or convenience of time zone.
C) The fact the students have been randomly assigned means we could use this data as a null distribution for a randomization test.
D) As we have observed both the final grades and the group students were in, this is an example of a cross-sectional study.

Question 12

Suppose a company wanted to understand how remote working was affecting their employees. In one of their staff surveys they had asked employees to rate their current sleep quality. 30 employees with generally poor sleep quality and 30 employees with generally excellent sleep quality were then invited to be part of a further study where they were asked whether or not they usually worked on their computer within 2 hours of their bedtime. The goal of the study was to understand if being exposed to the computer close to bedtime was associated with poor sleep quality. What kind of study is this?

A) Randomized control trial.
B) Prospective cohort study.
C) Retrospective cohort study.
D) Longitudinal study.
E) Case-control study.

Question 13

Suppose U of T is currently recruiting as study participants students graduating with one of the following degrees in 2020: Statistics Specialist, Data Science Specialist, Actuarial Science Specialist. The goal of the

study is to identify what students with these degrees are earning 5 years after graduating (2025) and whether there were any differences in their incomes by program. What kind of study is this?

A) Randomized control trial.
B) Prospective cohort study.
C) Retrospective cohort study.
D) Longitudinal study.
E) Case-control study.