

Stat 6021: Addressing Guided Question Set 2

Tom Lever

09/02/22

For this exercise, use the `.csv` data file that you created at the end of the previous guided question set, `new_students.csv`. As a reminder, the dataset contains information on students taking an introductory statistics class at a large public university. The columns of data are:

- * ``Student``: ID number on survey
- * ``Gender``: gender of student (male / female)
- * ``Smoke``: whether the student smokes (yes / no)
- * ``Marijuan``: whether the student smokes marijuana (yes / no)
- * ``DrivDrnk``: whether the student has ever driven while drunk (yes / no)
- * ``GPA``: student's current GPA
- * ``PartyNum``: number of days per month the student parties
- * ``DaysBeer``: number of days per month the student has at least two alcoholic drinks
- * ``StudyHrs``: number of hours spent studying per week
- * ``PartyAnimal``: whether the students parties more than 8 days per month (yes / no)
- * ``GPA.cat``: "low" if GPA is less than 3.0, "moderate" if GPA is at least 3.0 and less than 3.5, "high"

```
students_dataframe <-  
  read.csv(  
    file = "../Module_1--Data_Wrangling/Guided_Question_Set/new_students.csv"  
  )  
head(students_dataframe, n = 3)
```

```
## Student Gender Smoke Marijuan DrivDrnk GPA PartyNum DaysBeer StudyHrs  
## 1      1 female    No      Yes      Yes 3.40      4      6      7  
## 2      2 female    No      No       No 3.45      4      0     20  
## 3      3 male     No      No      Yes 3.89      9      4     30  
## PartyAnimal GPA.cat  
## 1           no moderate  
## 2           no moderate  
## 3          yes    high
```

1. Produce a frequency table of the number of students in each level of `GPA.cat`. If needed, be sure to arrange the order of the output appropriately. How many students are in each level of `GPA.cat`?

```
GPA_category_factor <-  
  factor(students_dataframe$GPA.cat, labels = c("low", "moderate", "high"))  
table(GPA_category_factor)
```

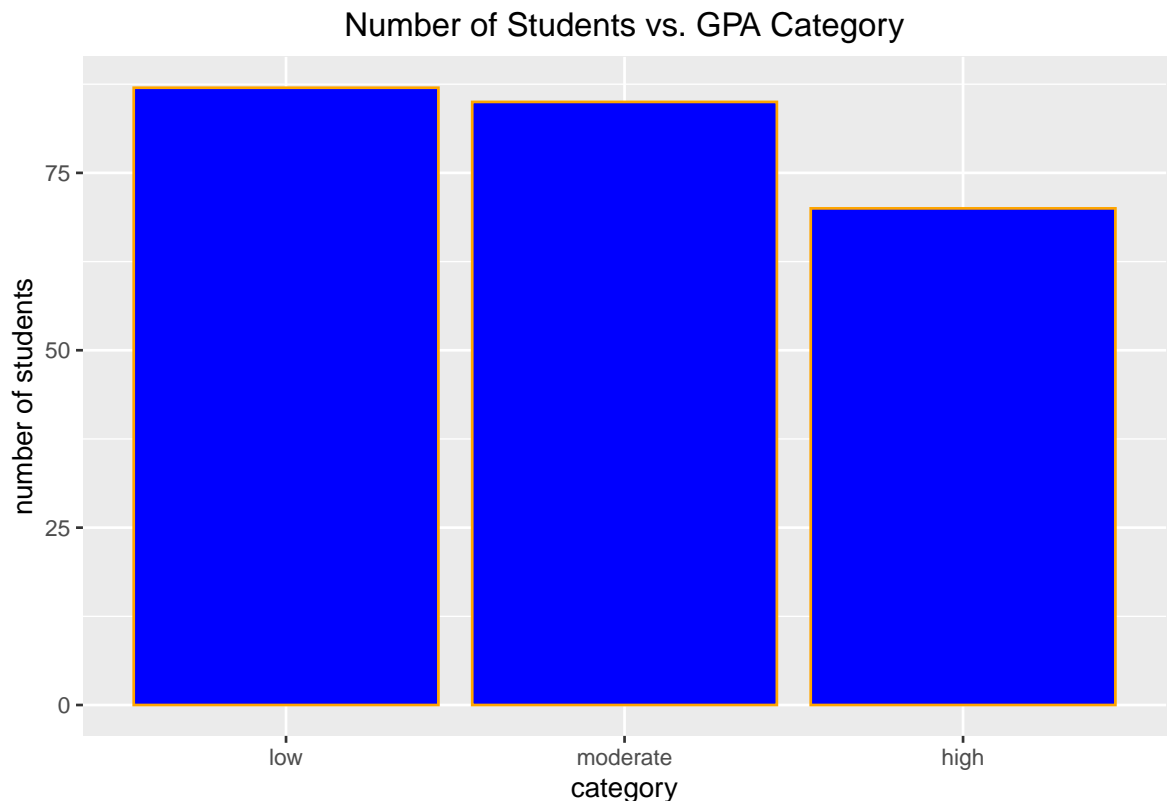
```
## GPA_category_factor  
##      low moderate    high  
##      70      87      85
```

For the students taking an introductory statistics class at a large public university, 70 have low GPA's, 87 have moderate GPA's, and 85 have high GPA's.

2. Produce a bar chart that summarizes the number of students in each level of `GPA.cat`. Be sure to add

appropriate labels and titles so that the bar chart conveys its message clearly to the reader. Be sure to remove the bar corresponding to the missing values.

```
library(dplyr)
library(ggplot2)
GPA_category <-
  students_dataframe %>%
    select(GPA.cat) %>%
    filter(!is.na(GPA.cat))
colnames(GPA_category) <- "GPA_category"
ggplot(data = GPA_category, aes(x = GPA_category)) +
  geom_bar(fill = "blue", color = "orange") +
  scale_x_discrete(limits = c("low", "moderate", "high")) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
  ) +
  labs(
    x = "category",
    y = "number of students",
    title = "Number of Students vs. GPA Category"
  )
)
```



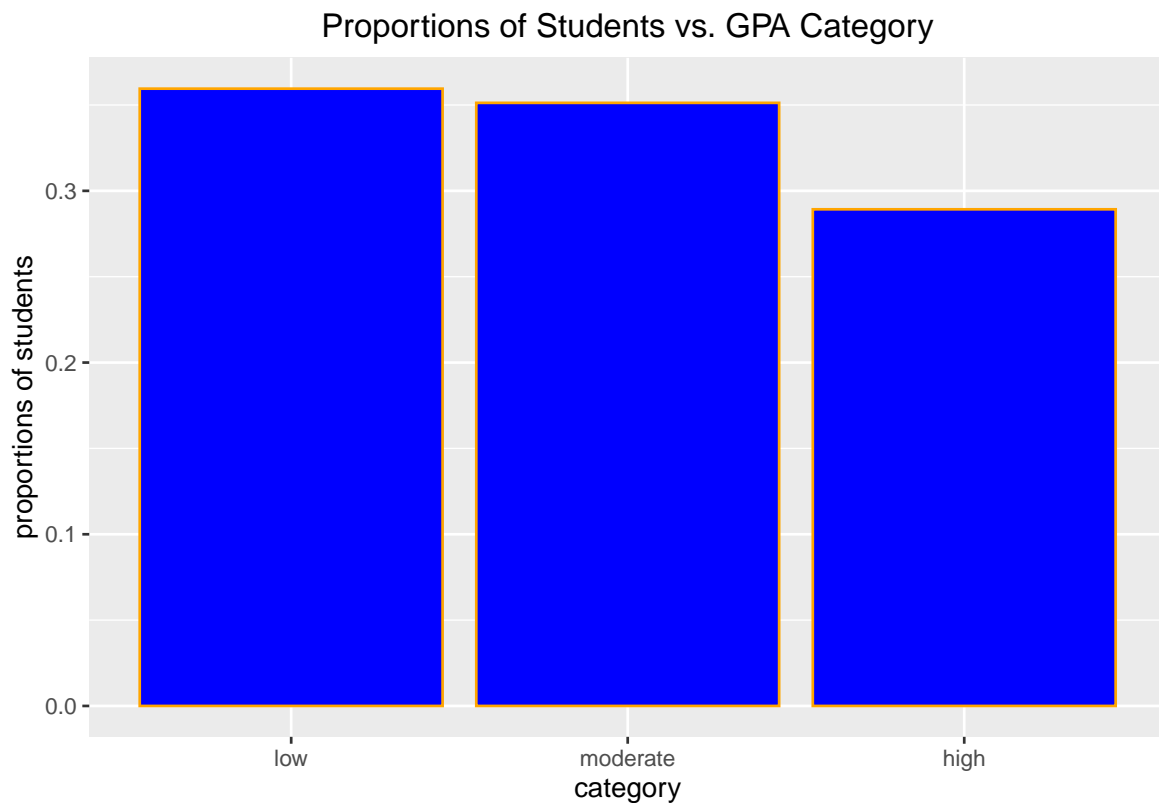
3. Create a similar bar chart as you did in Part 2, but with proportions instead of counts. Be sure to remove the bar corresponding to the missing values.

```
library(dplyr)
library(ggplot2)
number_of_students <- nrow(GPA_category)
```

```

proportion <-
  GPA_category %>%
    group_by(GPA_category) %>%
    summarize(numbers_of_students = n()) %>%
    mutate(proportions_of_students = numbers_of_students / number_of_students)
ggplot(data = proportion, aes(x = GPA_category, y = proportions_of_students)) +
  geom_bar(stat = "identity", fill = "blue", color = "orange") +
  scale_x_discrete(limits = c("low", "moderate", "high")) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
  ) +
  labs(
    x = "category",
    y = "proportions of students",
    title = "Proportions of Students vs. GPA Category"
  )

```



4. Produce a frequency table for the number of female and male students and the GPA category.

```

numbers_of_students_by_gender_and_GPA_category <-
  table(
    students_dataframe$Gender,
    factor(students_dataframe$GPA.cat, labels = c("low", "moderate", "high"))
  )
numbers_of_students_by_gender_and_GPA_category

```

```
##
```

```
##           low moderate high
##  female  46         41  52
##   male   24         46  33
```

```
chisq.test(numbers_of_students_by_gender_and_GPA_category)
```

```
##
##  Pearson's Chi-squared test
##
## data:  numbers_of_students_by_gender_and_GPA_category
## X-squared = 6.2312, df = 2, p-value = 0.04435
```

Given a significance level 0.05, since the above probability 0.04435 is less than the significance level, we reject the null hypothesis of the Pearson's Chi-squared test of independence, which states that there is no association between the row variable **Gender** and the column variable **GPA category**. We have sufficient evidence to conclude that there is an association between the row variable **Gender** and the column variable **GPA category**.

5. Produce a table for the percentage of GPA categories for each gender. For the percentages, round to 2 decimal places. Comment on the relationship between gender and GPA category.

```
percentages_of_students_by_gender_and_GPA_category <-
  round(prop.table(numbers_of_students_by_gender_and_GPA_category, 1) * 100, 2)
chisq.test(percentages_of_students_by_gender_and_GPA_category)
```

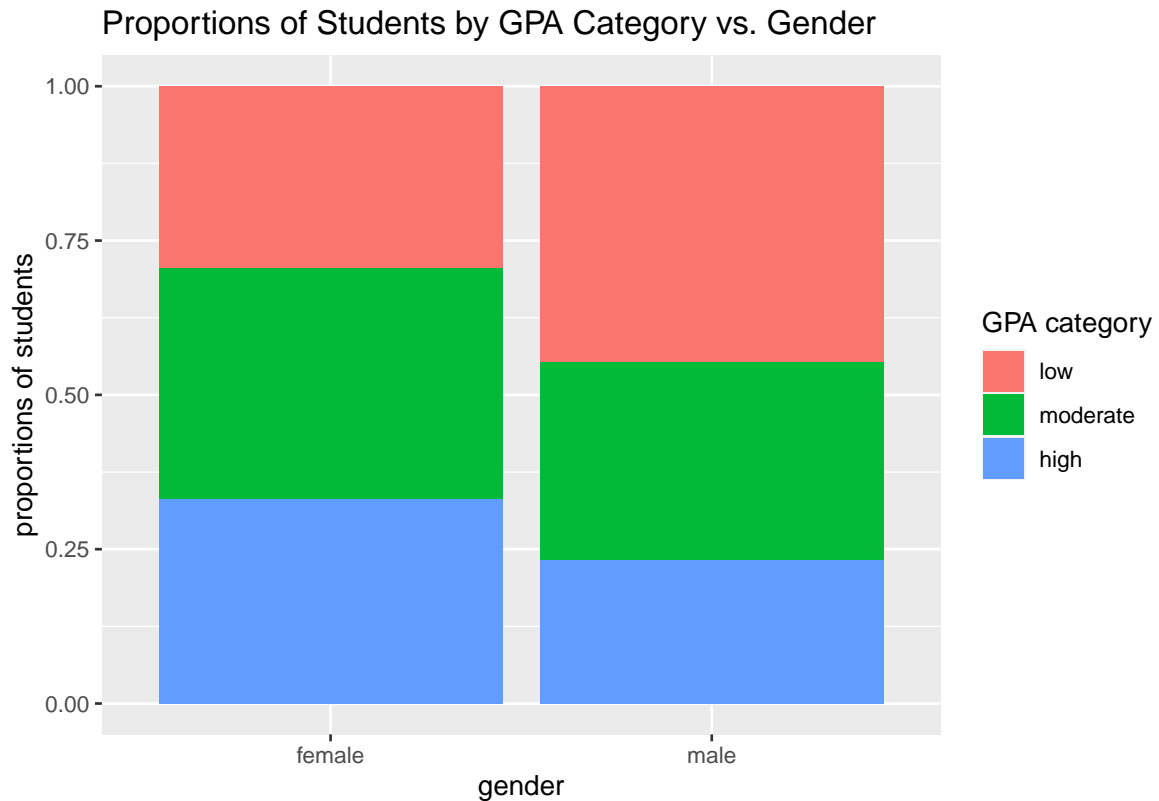
```
##
##  Pearson's Chi-squared test
##
## data:  percentages_of_students_by_gender_and_GPA_category
## X-squared = 5.2139, df = 2, p-value = 0.07376
```

Why are the results of these two Pearson's Chi-squared tests of independence different? See above comment.

6. Create a bar chart to explore the proportion of GPA categories for female and male students. Be sure to remove the bar corresponding to the missing values.

```
gender_and_GPA_category <-
  students_dataframe %>%
    select(Gender, GPA.cat) %>%
    filter(!is.na(GPA.cat))
colnames(gender_and_GPA_category) <- c("gender", "GPA_category")
proportions_by_gender_and_GPA_category <-
  gender_and_GPA_category %>%
    group_by(gender, GPA_category) %>%
    summarize(proportions_of_students = n()) %>%
    mutate(proportions_of_students = proportions_of_students / number_of_students)
ggplot(
  proportions_by_gender_and_GPA_category,
  aes(
    fill = factor(GPA_category, levels = c("low", "moderate", "high")),
    x = gender,
    y = proportions_of_students
  )
) +
  geom_bar(position="fill", stat="identity") +
  scale_fill_discrete(name = "GPA category") +
  labs(
```

```
x = "gender",
y = "proportions of students",
title = "Proportions of Students by GPA Category vs. Gender"
)
```



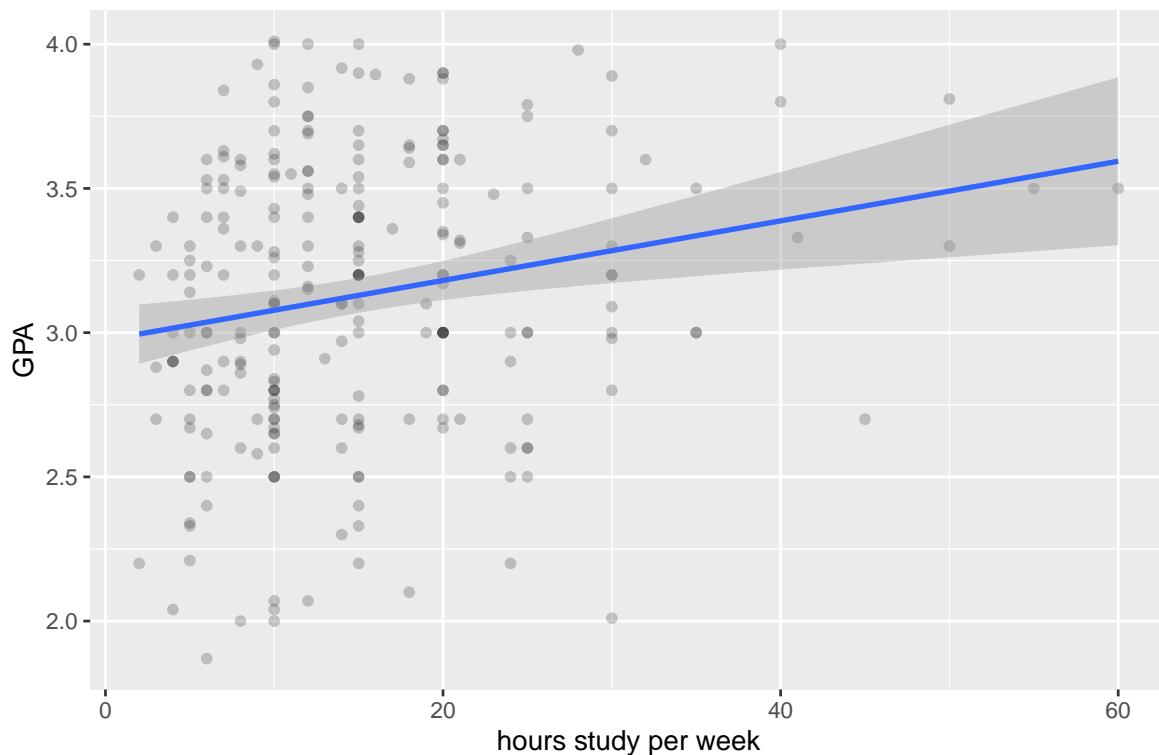
7. Create a bar chart similar to the bar chart in Part 6, but split by smoking status. Comment on this bar chart.
8. Create a scatterplot of GPA against the amount of hours spent studying per week. How would describe the relationship between GPA and amount of time spent studying?

```
hours_study_per_week_and_GPA <- students_dataframe %>% select(StudyHrs, GPA)
colnames(hours_study_per_week_and_GPA) <- c("hours_study_per_week", "GPA")
ggplot(hours_study_per_week_and_GPA, aes(x = hours_study_per_week, y = GPA)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x = "hours study per week",
    y = "GPA",
    title = "GPA vs. Hours Study Per Week"
  )
)
```

```
## Warning: Removed 7 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```

GPA vs. Hours Study Per Week



```
linear_model_for_GPA_versus_hours_study_per_week <-
  lm(GPA ~ hours_study_per_week, data = hours_study_per_week_and_GPA)
summary(linear_model_for_GPA_versus_hours_study_per_week)
```

```
##
## Call:
## lm(formula = GPA ~ hours_study_per_week, data = hours_study_per_week_and_GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27420 -0.33379  0.01744  0.39737  0.93237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.97435    0.05749  51.734 < 2e-16 ***
## hours_study_per_week 0.01033    0.00322   3.208  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4777 on 240 degrees of freedom
## (7 observations deleted due to missingness)
## Multiple R-squared:  0.04111,    Adjusted R-squared:  0.03712
## F-statistic: 10.29 on 1 and 240 DF,  p-value: 0.00152
```

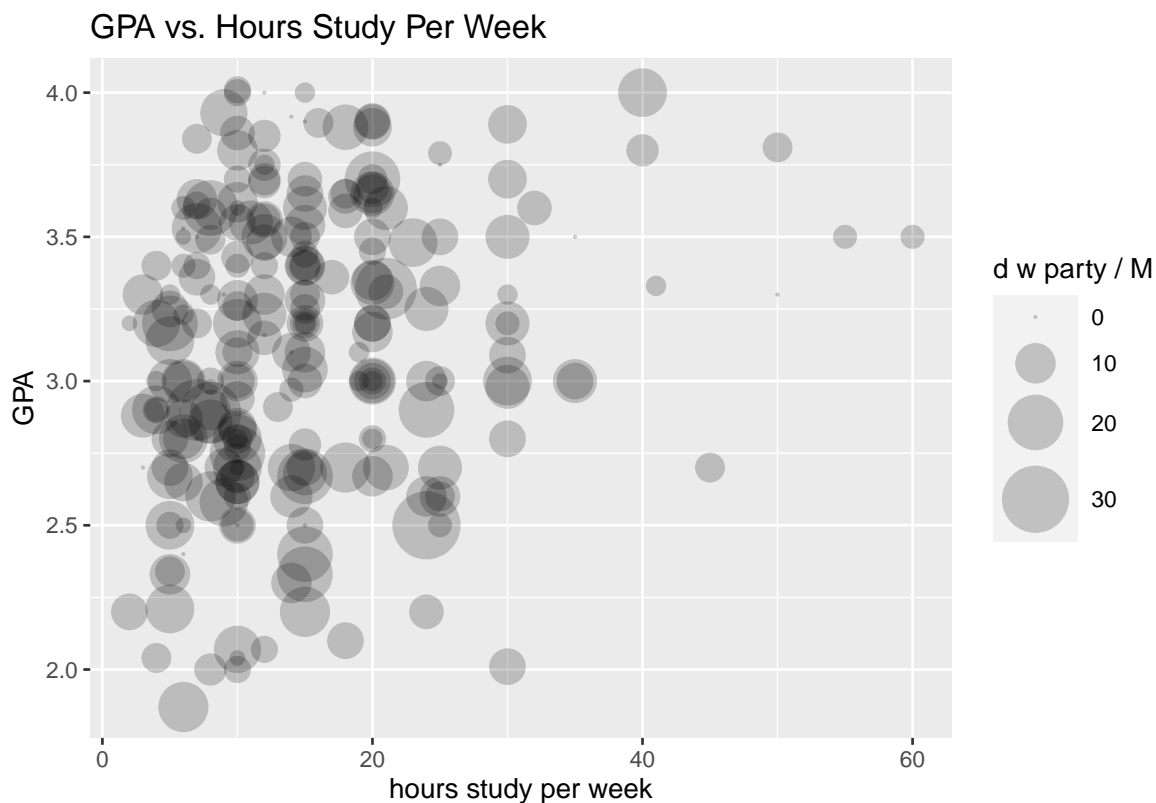
Given a significance level 0.05, since the above probability 0.002 is less than the significance level, we reject the null hypothesis of the linear regression t test, which states that there is no correlation between GPA and hours study per week. We have sufficient evidence to conclude that there is a correlation between GPA and hours study per week.

The proportion of the variation in GPA that is explained by the linear relationship between GPA and hours study per week is 0.037. Because this proportion is less than 0.8, this linear model is neither good for prediction nor precise.

9. Edit the scatterplot from Part 8 to include information about the number of days the student parties in a month.

```
hours_study_per_week_and_GPA <- students_dataframe %>% select(StudyHrs, GPA, PartyNum)
colnames(hours_study_per_week_and_GPA) <-
  c("hours_study_per_week", "GPA", "number_of_days_with_party_per_month")
ggplot(
  hours_study_per_week_and_GPA,
  aes(
    x = hours_study_per_week,
    y = GPA,
    size = number_of_days_with_party_per_month
  )
) +
  geom_point(alpha = 0.2) +
  labs(
    x = "hours study per week",
    y = "GPA",
    size = "d w party / M",
    title = "GPA vs. Hours Study Per Week"
  ) +
  scale_size(range = c(0.1, 12))
```

Warning: Removed 12 rows containing missing values (geom_point).



10. Edit the scatterplot from Part 9 to include information about whether the student smokes or not.

```
hours_study_per_week_and_GPA <- students_dataframe %>% select(StudyHrs, GPA, PartyNum, Smoke)
colnames(hours_study_per_week_and_GPA) <-
  c("hours_study_per_week", "GPA", "number_of_days_with_party_per_month", "smokes")
ggplot(
  hours_study_per_week_and_GPA,
  aes(
    x = hours_study_per_week,
    y = GPA,
    size = number_of_days_with_party_per_month,
    color = smokes
  )
) +
  geom_point(alpha = 0.2) +
  labs(
    x = "hours study per week",
    y = "GPA",
    size = "d w party / M",
    title = "GPA vs. Hours Study Per Week"
  ) +
  scale_size(range = c(0.1, 12))
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

