Guided Set 2

Michael Puchalski

2025-01-28

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df<-read.csv("new_students.csv", header = TRUE)
```

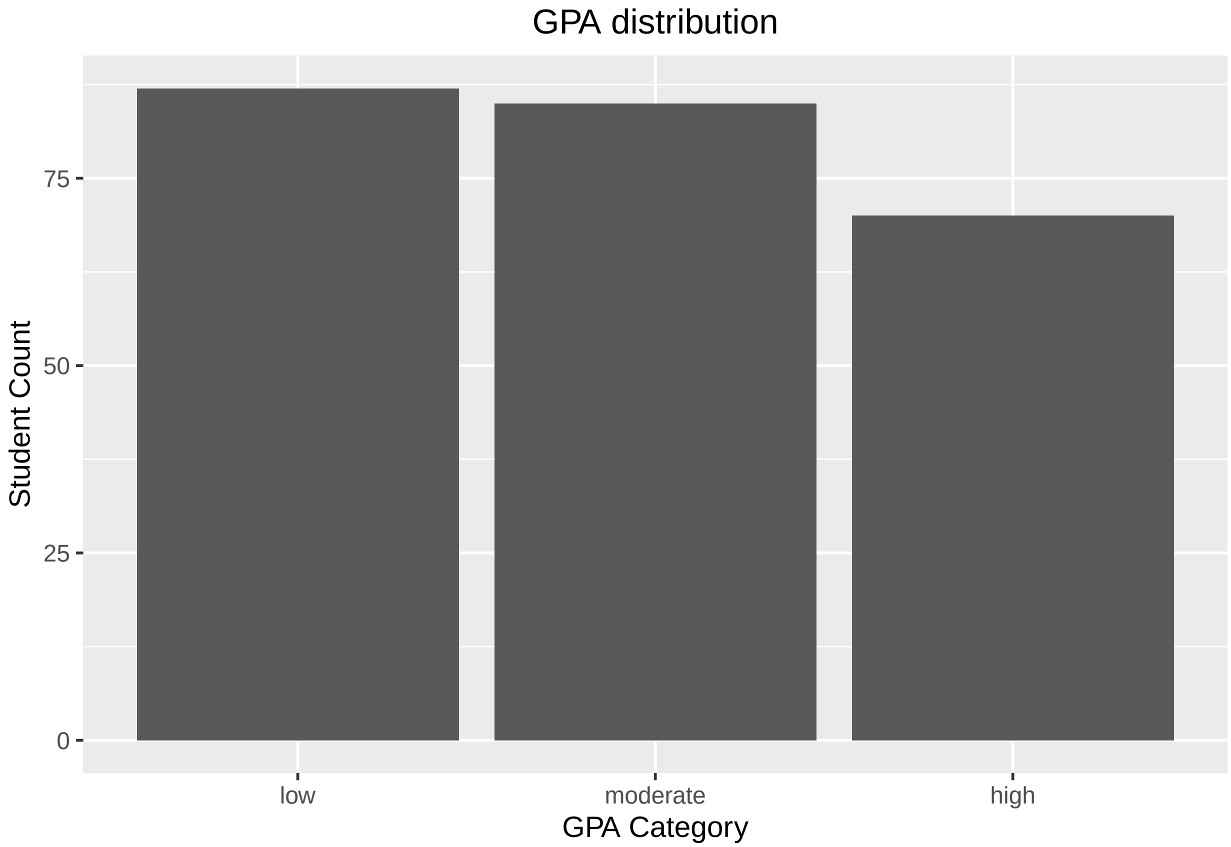
#Question 1 Frequency Table of the number of students in each level of GPA.cat then arrange if needed

```
df$GPA.cat<- factor(df$GPA.cat, levels=c("low","moderate","high"))
# levels(df$GPA.cat) checks the order of levels
table(df$GPA.cat)
```

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

#Question 2 Create a bar chart for this data

```
df_clean<-df%>%
  drop_na(GPA.cat)
ggplot(df_clean, aes(x=GPA.cat))+
  geom_bar() +
  theme(axis.text.x = element_text(angle=0),
        plot.title = element_text(hjust=0.5))+
  labs(x="GPA Category", y = "Student Count", title = "GPA distribution")
```

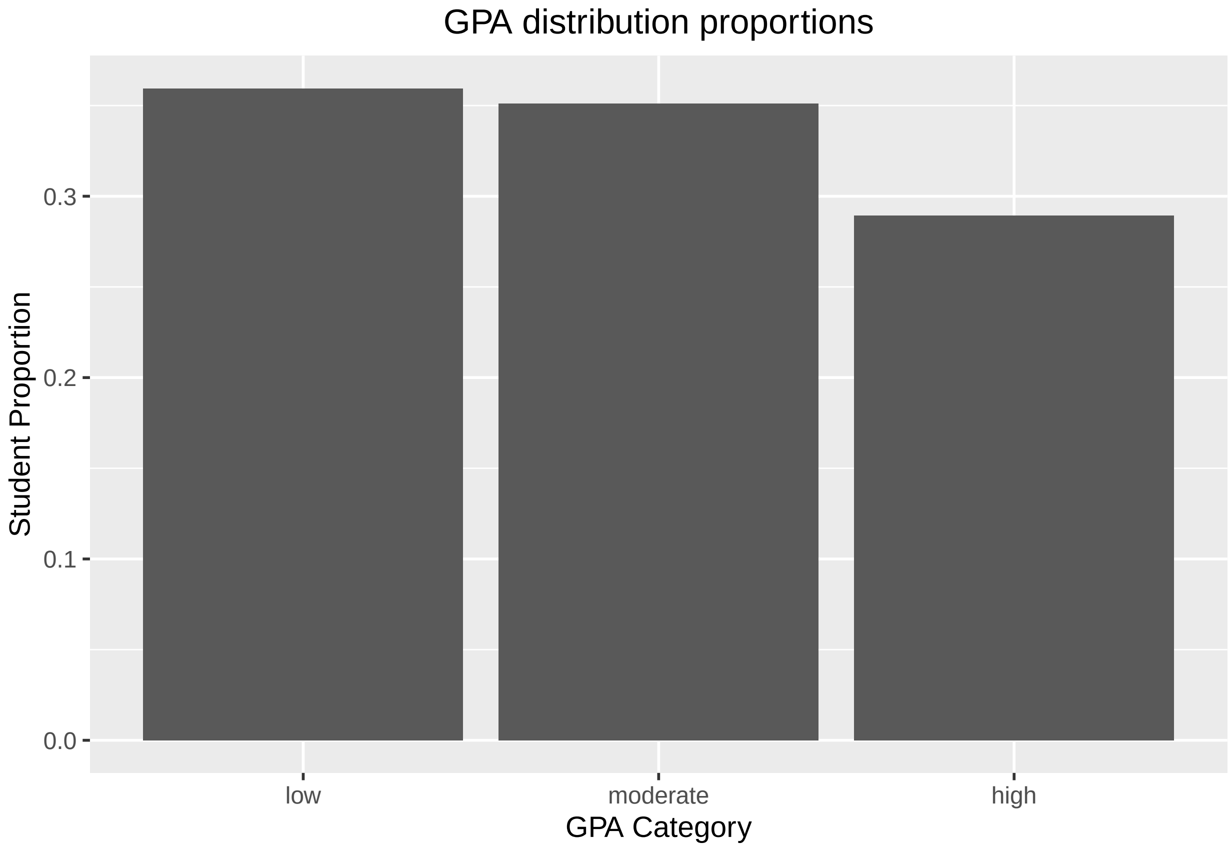


#Question 3 Create a bar chart with proportions

```
prop_df<-df_clean%>%
  group_by(GPA.cat)%>%
  summarize(Counts=n())%>%
  mutate(Percent = Counts/nrow(df_clean))
prop_df
```

```
## # A tibble: 3 x 3
##   GPA.cat  Counts Percent
##   <fct>     <int>   <dbl>
## 1 low          87   0.360
## 2 moderate     85   0.351
## 3 high         70   0.289
```

```
ggplot(prop_df, aes(x=GPA.cat, y=Percent))+
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 0),
        plot.title = element_text(hjust = 0.5))+
  labs(x="GPA Category", y = "Student Proportion", title = "GPA distribution proportions")
```



#Question 4 Two Way table for the number of male and female students and the GPA category

```
two_way_table<-table(df_clean$Gender, df_clean$GPA.cat)
two_way_table
```

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

#Question 5 Produce a percentage table for the proportion of GPA categories

```
round(prop.table(two_way_table, 1)*100, 2)
```

```
##
##             low moderate  high
##   female 29.50    37.41 33.09
##   male   44.66    32.04 23.30
```

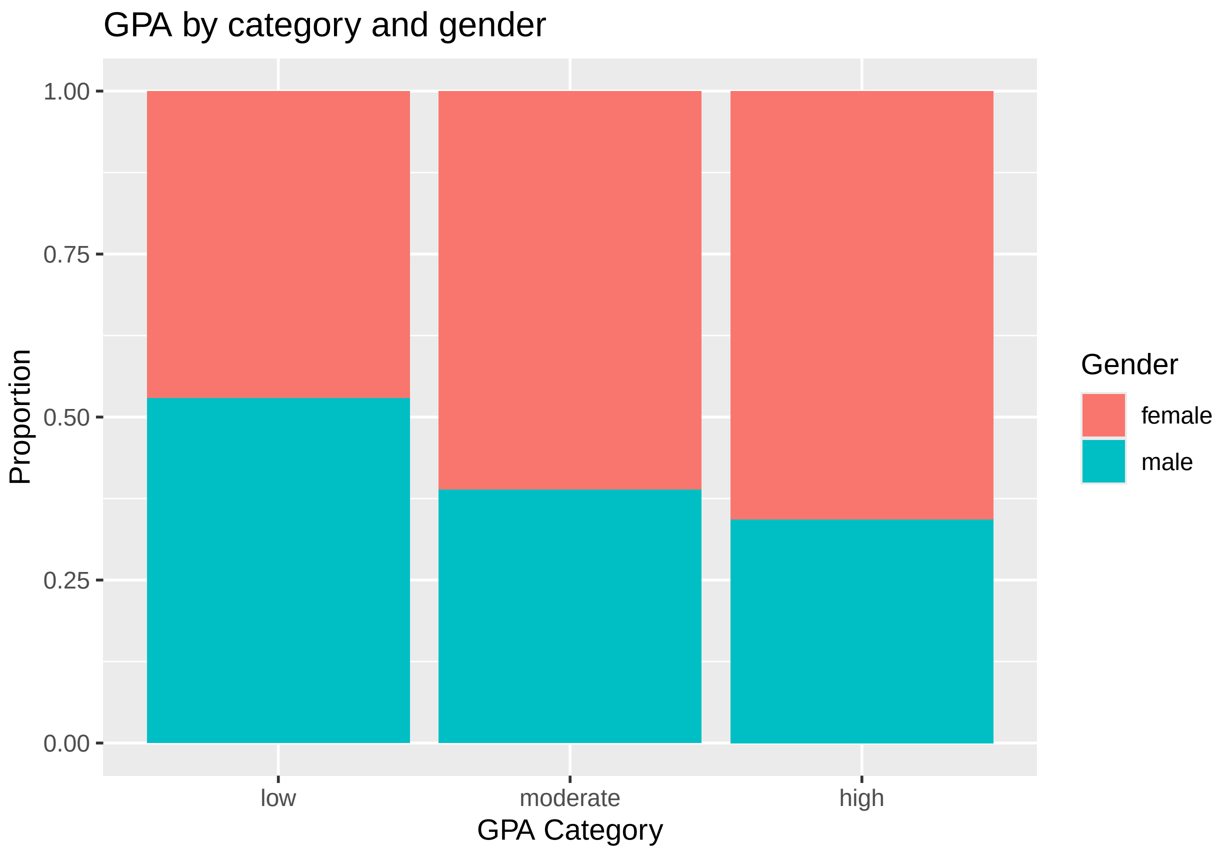
```
round(prop.table(two_way_table, 2)*100, 2)
```

```
##
##             low moderate  high
##   female 47.13    61.18 65.71
##   male   52.87    38.82 34.29
```

According to these tables it shows that females make up a larger percentage of the moderate and high GPA.categories with men making up just over half of the low GPA category. Along with this, men show to have a higher grouping in the low vs high end where the opposite is true for females.

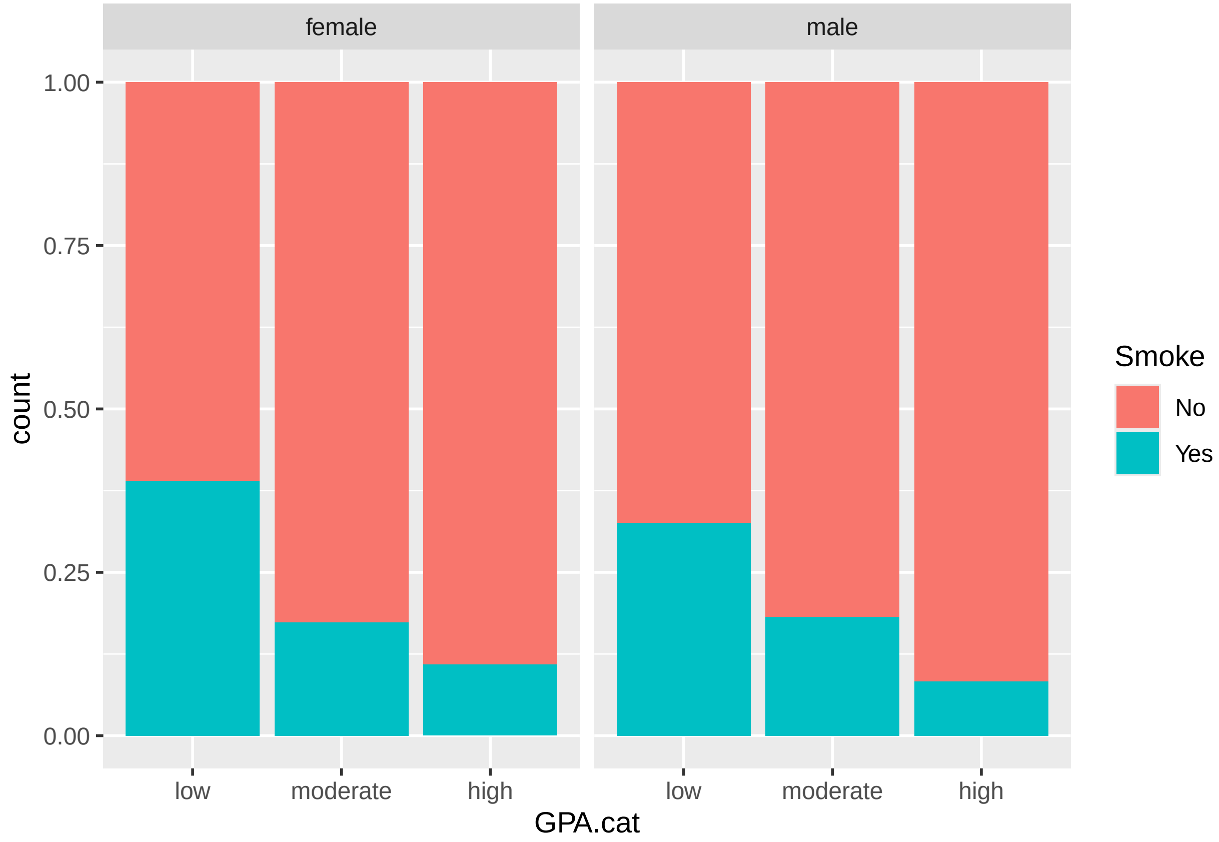
#Question 6 make a bivariute bar chart

```
ggplot(df_clean, aes(x = GPA.cat, fill = Gender))+
  geom_bar(position = "fill")+
  labs(x="GPA Category", y = "Proportion", title = "GPA by category and gender")
```



#Question 7 Split this further by smoker vs nonsmoker

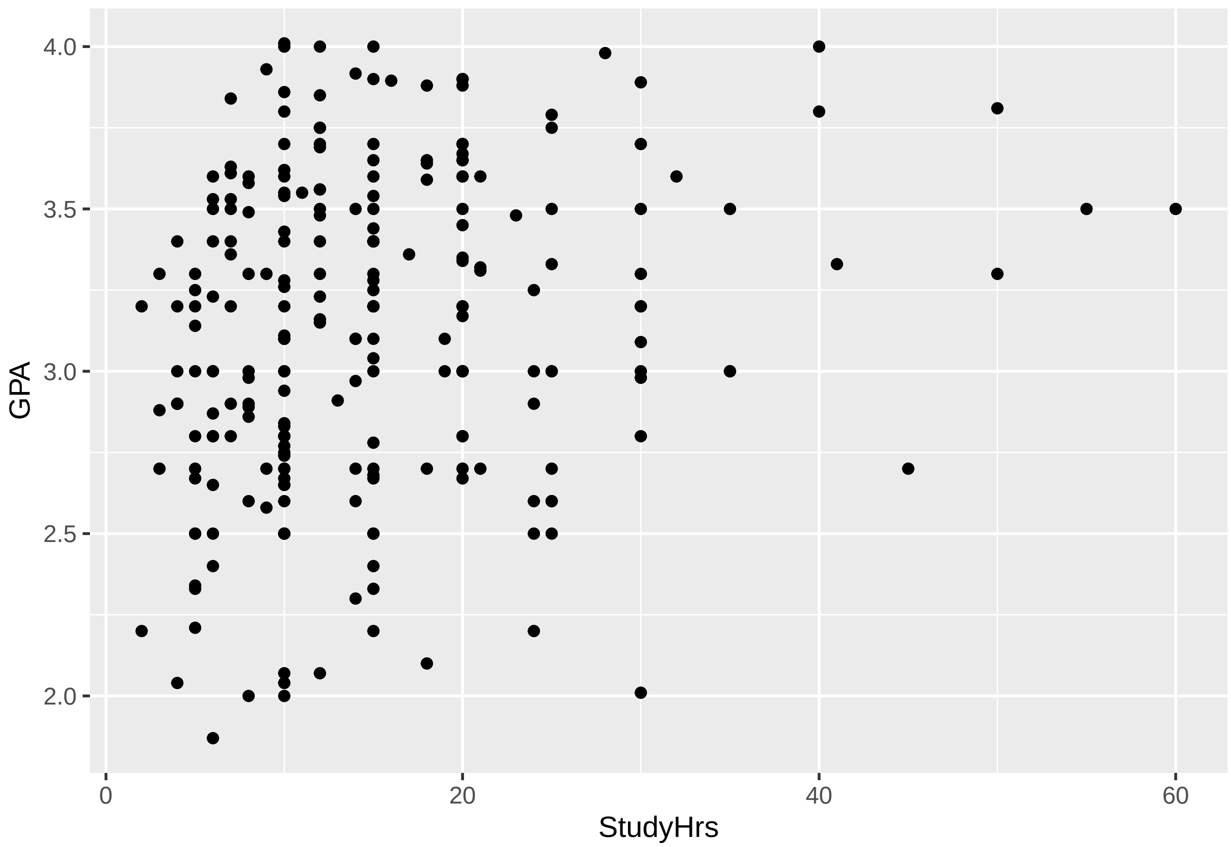
```
ggplot(df_clean, aes(x = GPA.cat, fill = Smoke))+
  geom_bar(position = "fill")+
  facet_wrap(~Gender)
```



This chart would probably be better as a count chart, but the way to give the most information is by allowing it to wrap by gender which shows that the more successful students both male and female have a lower rate of being smokers than in the lower GPA categories.That being said, it is a chart that needs context to allow it to be a proper visualization.

#Question 8 Creat GPA vs study hours scatter plot

```
ggplot(df_clean, aes(x = StudyHrs, y = GPA))+
  geom_point()
```

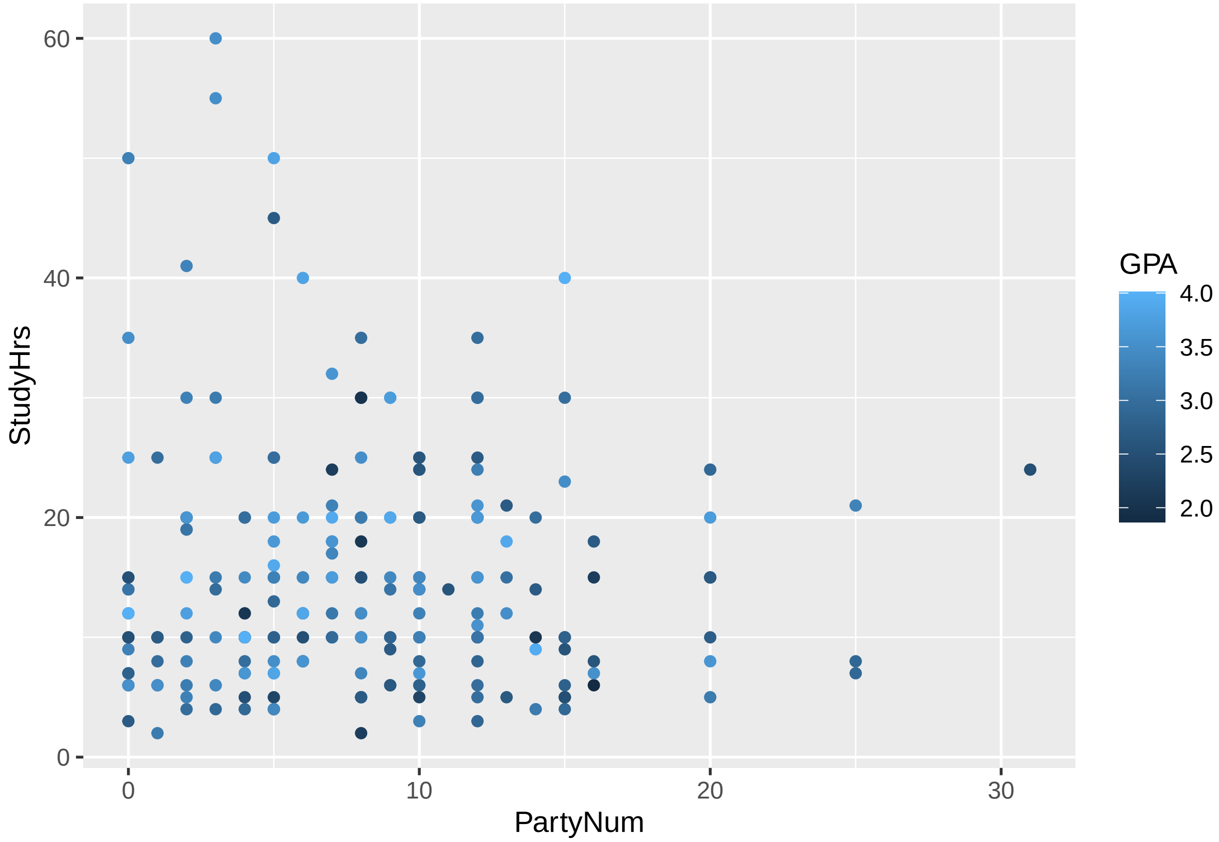


This chart shows that there is no guarantee of success, but that when more than 30 hours are spent studying, the lowest GPA is a 2.75. So, while increasing study hours does not guarantee an increase in GPA, studying more can correlate. Along with this, it shows that su=ome students can succeed with relatively low study hrs per week.

#Question 9 Make the scatter plot include party days too

```
ggplot(df_clean, aes(x = PartyNum, y = StudyHrs, color = GPA))+
  geom_point()
```

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

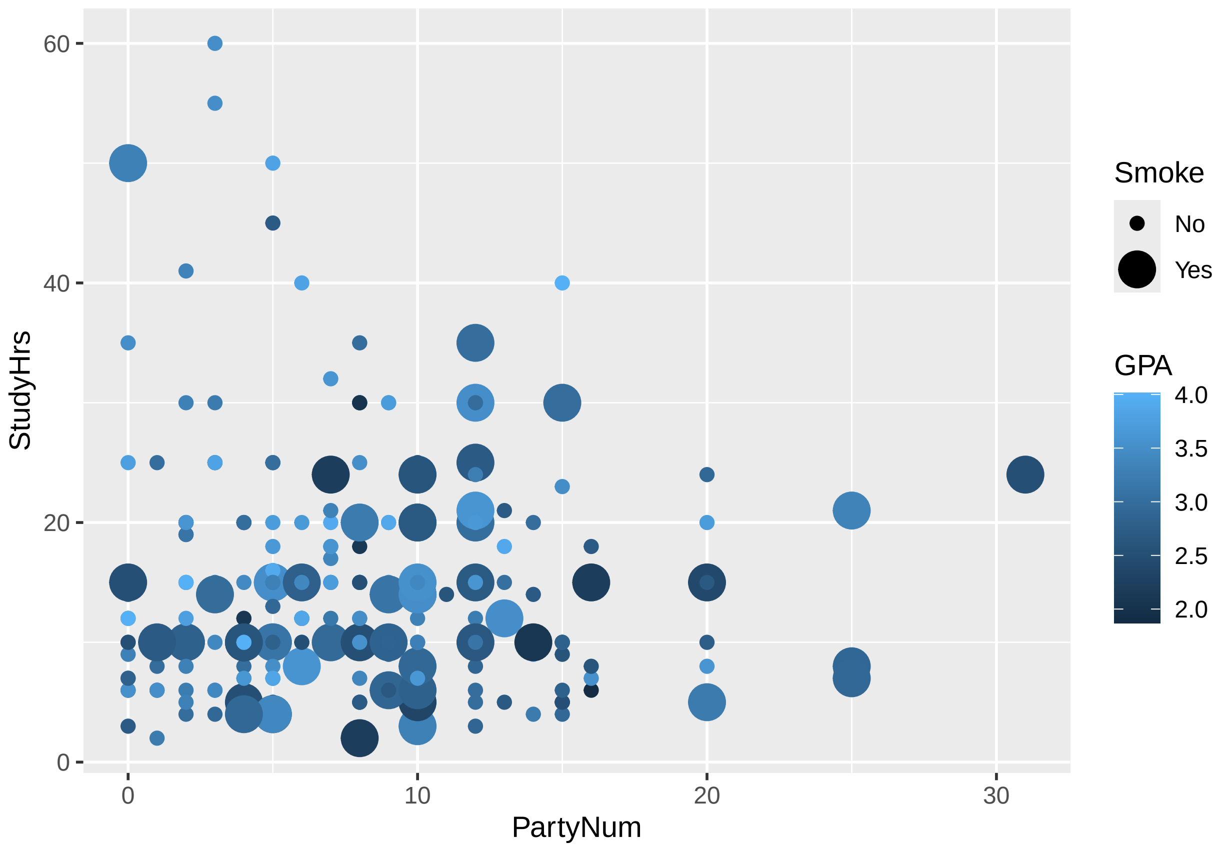


#Question 10 same thing plus smoking

```
ggplot(df_clean, aes(x = PartyNum, y = StudyHrs, color = GPA, size = Smoke))+
  geom_point()
```

```
## Warning: Using size for a discrete variable is not advised.
```

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


StudyHrs
60
40
20
0
0
10
20
30
PartyNum
Smoke
No
Yes
GPA
4.0
3.5
3.0
2.5
2.0