Solutions to Guided Question Set 2

```
library(tidyverse)

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

Question 1

```
table(students.df$GPA.cat)
```

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

```
##reorder levels of GPA.cat
students.df$GPA.cat<-factor(students.df$GPA.cat, levels=c("low","moderate","high"))
levels(students.df$GPA.cat)
```

```
## [1] "low"      "moderate" "high"
```

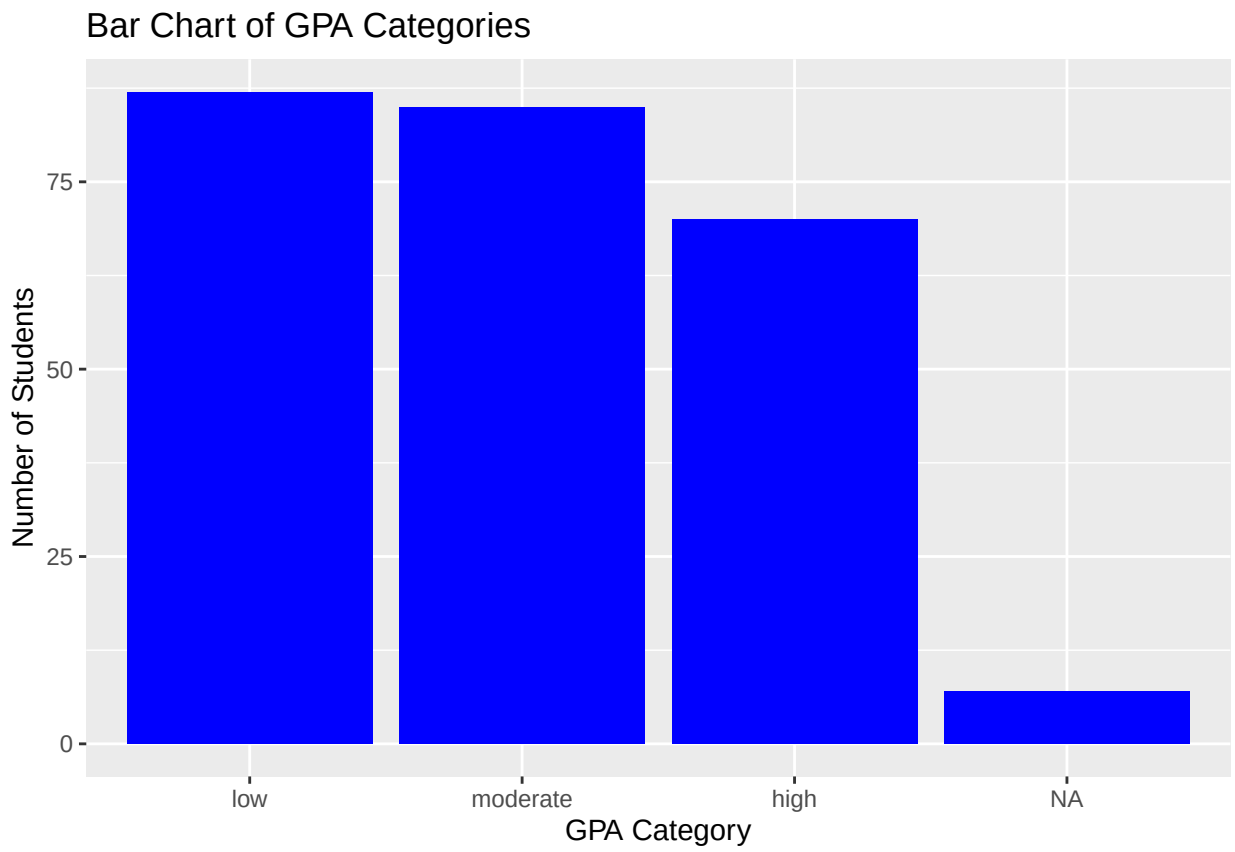
```
##recreate table with proper order
table(students.df$GPA.cat)
```

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

Notice that the original table should be reordered to make more sense in this context. We have 87 students who have low GPAs, 85 with moderate GPAs, and 70 with high GPAs.

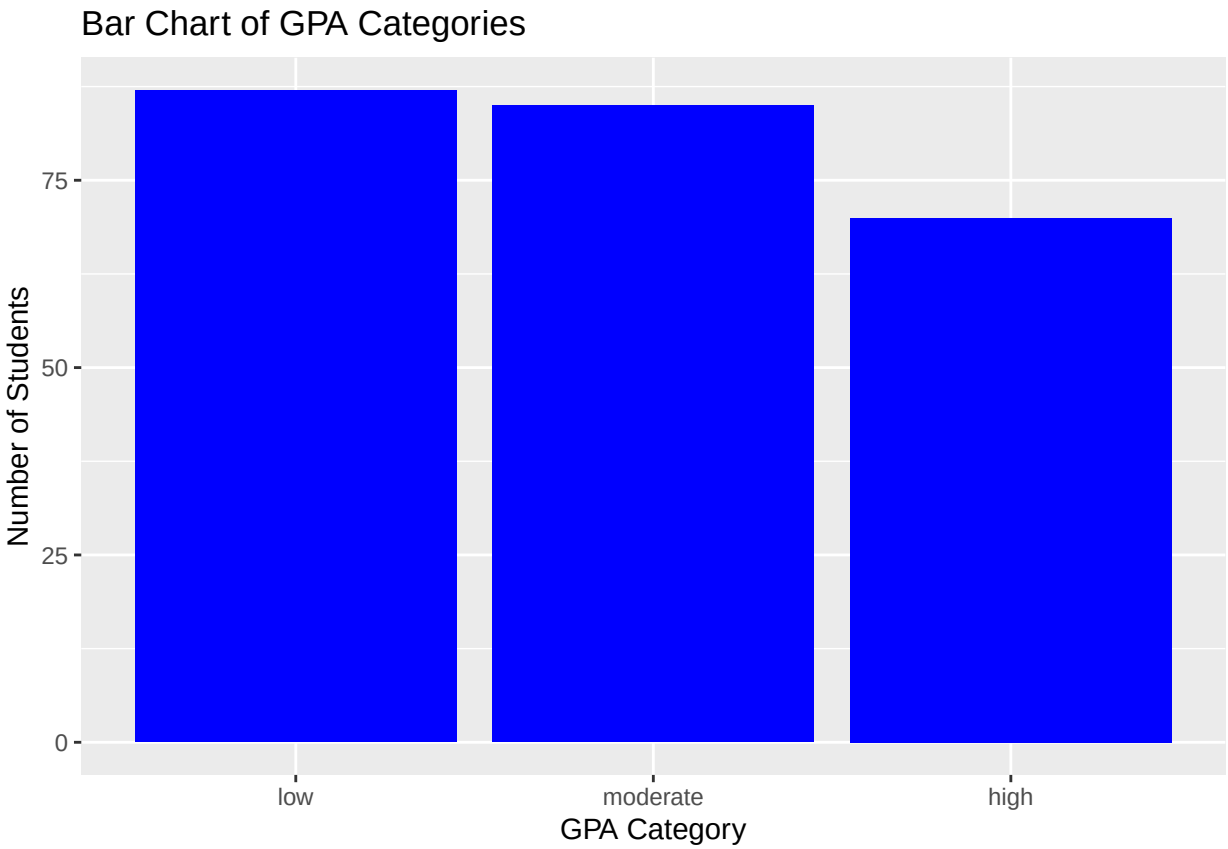
Question 2

```
ggplot(students.df, aes(x=GPA.cat))+
  geom_bar(fill="blue")+
  labs(x="GPA Category", y="Number of Students",
       title="Bar Chart of GPA Categories")
```



Notice that there is a bar for students with missing value for GPA category. To remove this bar

```
students.df %>%
  filter(!is.na(GPA.cat)) %>%
  ggplot(aes(x=GPA.cat))+
  geom_bar(fill="blue")+
  labs(x="GPA Category", y="Number of Students",
       title="Bar Chart of GPA Categories")
```

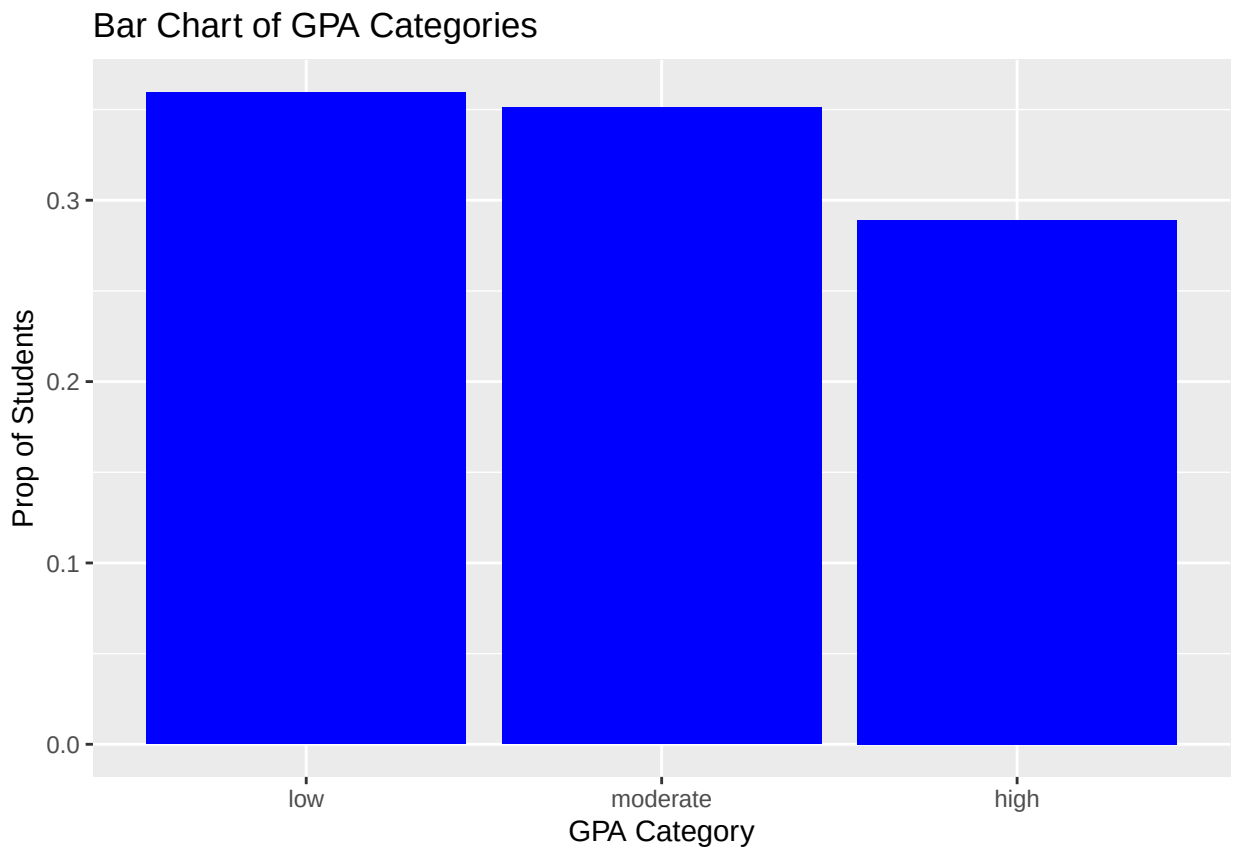


Question 3

To create a bar chart of proportions for GPA category

```
##Create bar chart with prop instead of counts
newData<-students.df%>%
  filter(!is.na(GPA.cat)) %>%
  group_by(GPA.cat)%>%
  summarize(Counts=n())%>%
  mutate(Percent=Counts/sum(Counts))

newData %>%
  ggplot(aes(x=GPA.cat, y=Percent))+
  geom_bar(fill="blue", stat="identity")+
  labs(x="GPA Category", y="Prop of Students",
       title="Bar Chart of GPA Categories")
```



Question 4

```
mytab<-table(students.df$Gender,students.df$GPA.cat)
mytab
```

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

Question 5

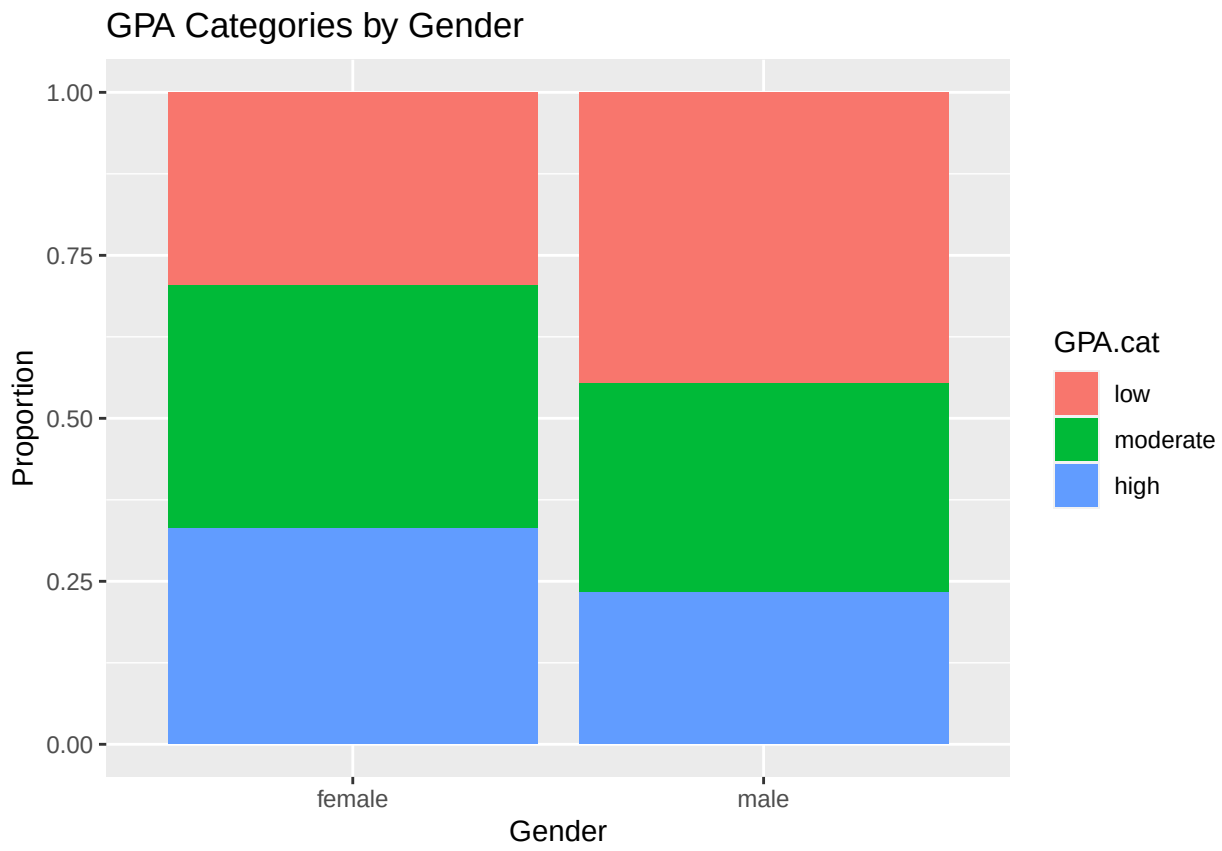
```
##prop of GPA.cat for each gender
round(prop.table(mytab,1)*100, 2)
```

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

A higher proportion of female students have high GPAs compared to male students (33.09% vs 23.30%). Not surprisingly, a lower proportion of female students have low GPAs compared to male students (29.50% vs 44.66%). The proportion of female and male students with moderate GPAs are about the same. Overall, female students are more likely to have high GPAs and less likely to have low GPAs than male students.

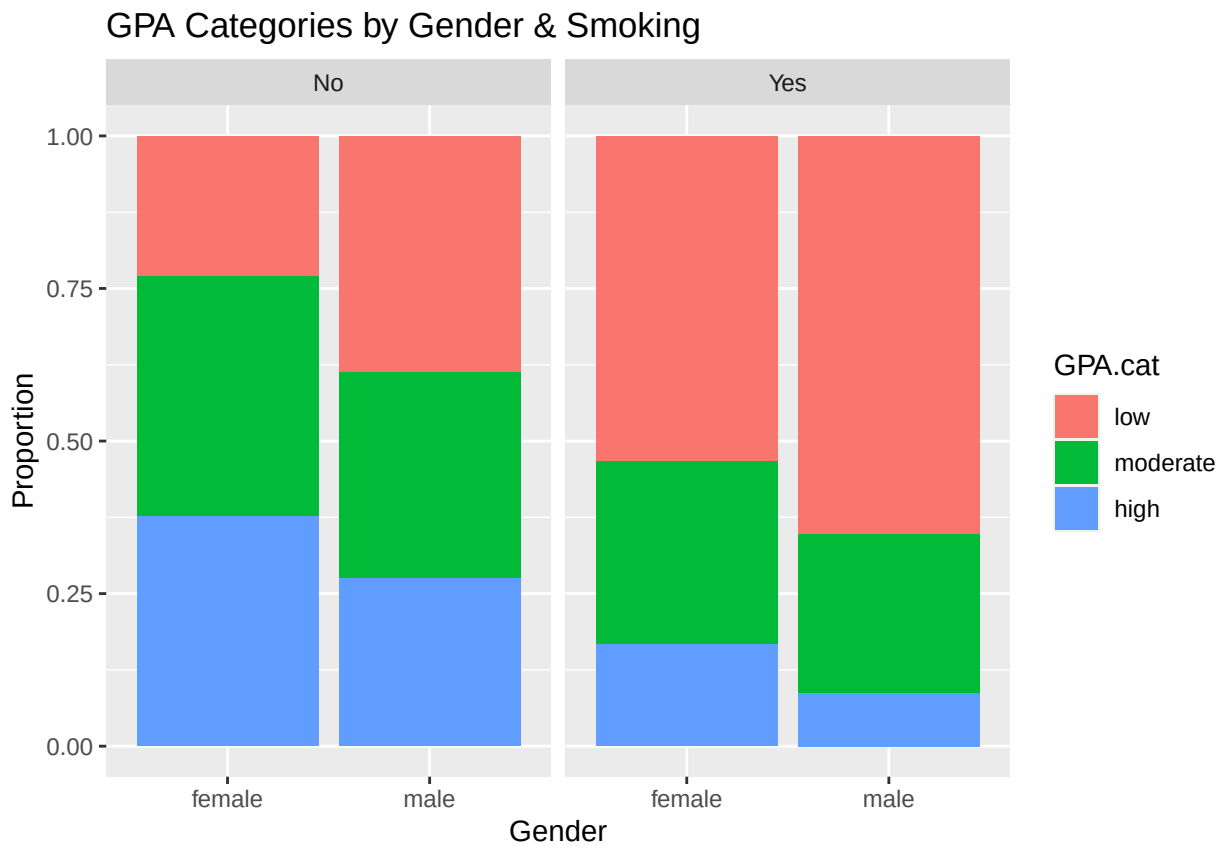
Question 6

```
students.df %>%
  filter(!is.na(GPA.cat)) %>%
  ggplot(aes(x=Gender, fill=GPA.cat))+
  geom_bar(position = "fill")+
  labs(x="Gender", y="Proportion",
       title="GPA Categories by Gender")
```



Question 7

```
students.df %>%
  filter(!is.na(GPA.cat)) %>%
  ggplot(aes(x=Gender, fill=GPA.cat))+
  geom_bar(position = "fill")+
  facet_wrap(~Smoke)+
  labs(x="Gender", y="Proportion",
       title="GPA Categories by Gender & Smoking")
```

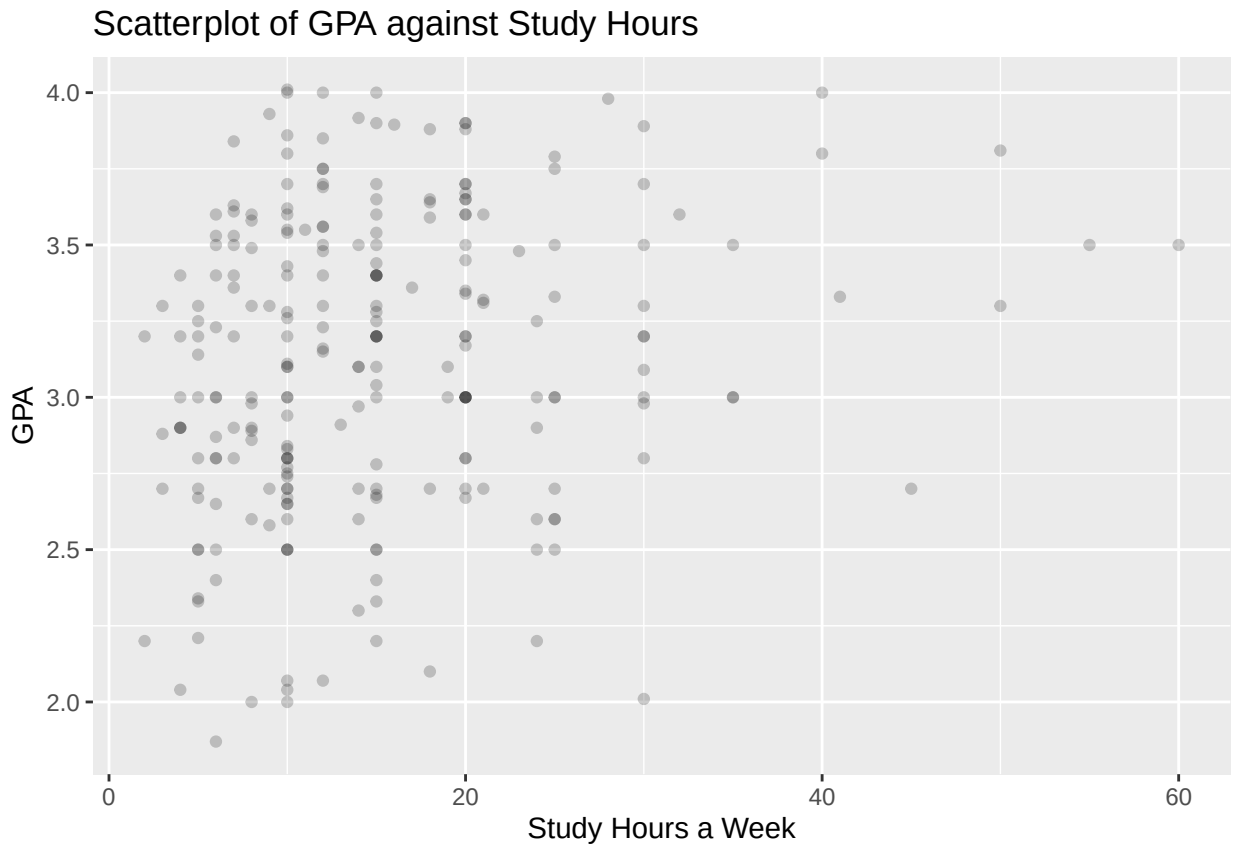


We see the same overall trend from the previous two questions: female students are more likely to have high GPAs, and less likely to have low GPAs, than male students, regardless of smoking status. When comparing the bar charts between smokers and non smokers, we can see that smokers are more likely to have low GPAs, and less likely to have high GPAs compared to non-smokers.

Question 8

```
ggplot(students.df, aes(x=StudyHrs,y=GPA))+
  geom_point(alpha=0.2)+
  labs(x="Study Hours a Week", y="GPA",
       title="Scatterplot of GPA against Study Hours")
```

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

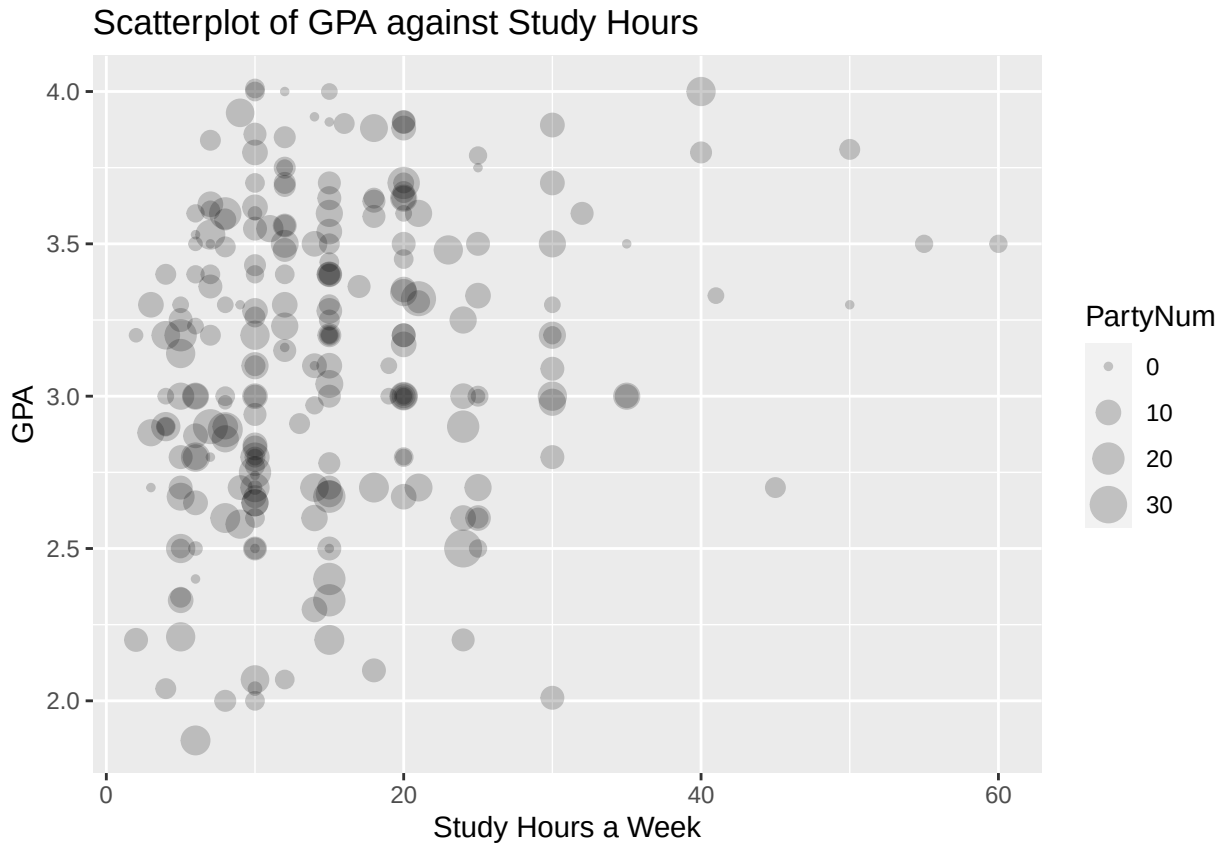


There appears to be some relationship between GPA and the amount of time spent studying. Generally, the more time spent studying, the higher the GPA, although this relationship is not very strong. The absence of data points in the bottom right quadrant does inform us that students who study a lot (more than 40 hours) almost always have a GPA higher than 3.25.

Question 9

```
ggplot(students.df, aes(x=StudyHrs,y=GPA,
                        size=PartyNum))+
  geom_point(alpha=0.2)+
  labs(x="Study Hours a Week", y="GPA",
       title="Scatterplot of GPA against Study Hours")
```

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



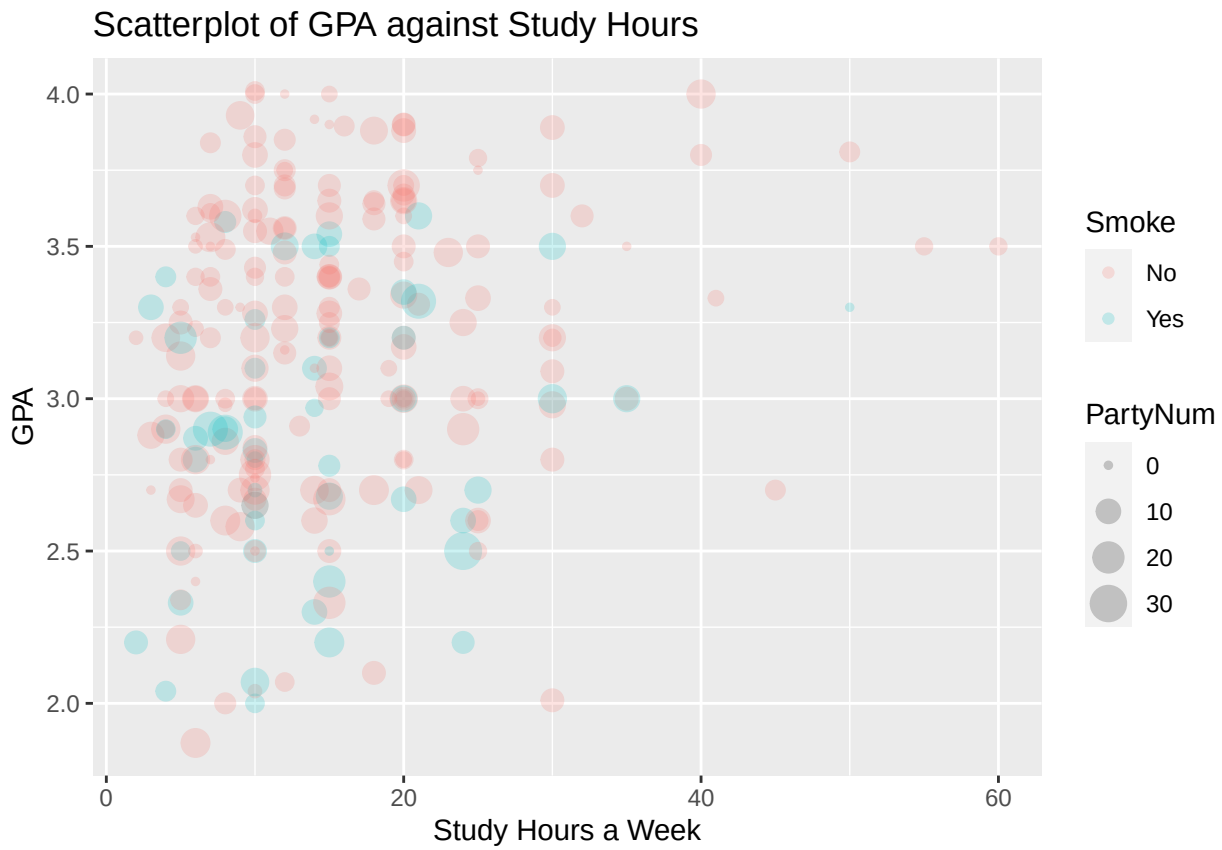
Looking at the top right quadrant, we see an individual who parties between 20 and 30 times a week, but studies 40 hours a week and has a 4.0 GPA. The rest of the students in this quadrant party between 10 and 20 times a month.

The picture in the left half of the plot is less clear, we see plots of all sizes that seem to be randomly scattered with little apparent pattern.

Question 10

```
ggplot(students.df,aes(x=StudyHrs,y=GPA,
                       size=PartyNum,
                       color=Smoke))+
  geom_point(alpha=0.2)+
  labs(x="Study Hours a Week", y="GPA",
       title="Scatterplot of GPA against Study Hours")
```

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



The red plots represent non-smokers, and they seem to have higher GPAs than the blue plots (smokers).

Students who study more than 30 hours a week tend to be non-smokers.

Note: with visualizations, there may be other ways of providing the needed information requested. It may take some trial and error to see what specific visualization works best for a particular question.