Solutions to Guided Question Set 1

Question 1

The first column is just an ID number of the student in the survey, and is just a label of little interest to any analysis, so it should be removed from the data. `students.df` is the data frame without this column.

```
students<-read.table("students.txt", header=TRUE)
students.df<-students[,-1]
```

Question 2

```
##number of students
nrow(students.df)
```

```
## [1] 249
```

There are 249 students in this data set.

Question 3

```
missing<-students.df[!complete.cases(students.df),]
nrow(missing)
```

```
## [1] 12
```

There are 12 students with missing values for some variables in this data set.

Question 4

```
apply(students.df[,c(5:8)],2,median,na.rm=T)
```

```
##      GPA PartyNum DaysBeer StudyHrs
##      3.2      8.0      8.0     14.0
```

The median GPA is 3.2. The median number of days partying a month is 8.0. The median number of days drinking at least two alcholic drinks a month is 8.0. The median number of hours studying a week is 14.0 hours.

Question 5

```
##mean of studyhrs by gender
tapply(students.df$StudyHrs, students.df$Gender, mean, na.rm=T)
```

```
##   female     male
## 15.40690 14.70192
```

```
##SD of studyhrs by gender
tapply(students.df$StudyHrs, students.df$Gender, sd, na.rm=T)
```

```
##    female      male
##  8.972564 10.198877
```

The mean study hours are 15.41 for females, and 14.70 for males. The mean is slightly higher for female students.

The standard deviation of study hours is 8.97 for females, and 10.20 for males. These values are fairly large, compared to the means, indicating a wide spread of time spent studying among the students, moreso for male.

Question 6

```
##mean of studyhrs by gender
study.mean<-tapply(students.df$StudyHrs, students.df$Gender, mean, na.rm=T)

##SD of studyhrs by gender
study.sd<-tapply(students.df$StudyHrs, students.df$Gender, sd, na.rm=T)
```

```
##sample sizes of genders
sample.size<-table(students.df$Gender)

##t multipliers
t.female<-qt(0.975,sample.size[1]-1)
t.male<-qt(0.975,sample.size[2]-1)

##lower and upper bound for CI, female
study.mean[1]-t.female*study.sd[1]/sqrt(sample.size[1])
```

```
##   female
## 13.93409
```

```
study.mean[1]+t.female*study.sd[1]/sqrt(sample.size[1])
```

```
##  female
## 16.8797
```

```
##lower and upper bound for CI, male
study.mean[2]-t.male*study.sd[2]/sqrt(sample.size[2])
```

```
##    male
## 12.7185
```

```
study.mean[2]+t.male*study.sd[2]/sqrt(sample.size[2])
```

```
##     male
## 16.68535
```

The 95% CI of study hours for female students is (13.934, 16.879) hours per week.

The 95% CI of study hours for male students is (12.718, 16.685) hours per week.

Since the CIs overlap with each other, we do not have evidence the mean study hours per week differs between female and male students.

Question 7

```
##median studyhrs by gender and smoke
tapply(students.df$StudyHrs, list(students.df$Gender, students.df$Smoke),
       median, na.rm=T)
```

```
##        No Yes
## female 15  10
## male   12  14
```

Female students who smoke have a median of 10 hours spent studying per week, which is lower for female students who do not smoke: 15 hours per week.

For male students, we note the opposite relationship: male students who smoke have a higher median than male students who do not smoke, 14 hours compared to 12 hours per week.

This observations indicates the possibility of an interaction between smoking and gender: the effect of smoking on study hours depends on gender.

Question 8

```
##new variable PartyAnimal
PartyAnimal<-ifelse(students.df$PartyNum>8, "yes", "no")
```

Question 9

```
##new variable GPA.cat
GPA.cat<-cut(students.df$GPA,
             breaks = c(-Inf, 3.0, 3.5, Inf),
             right= FALSE,
             labels = c("low", "moderate", "high"))
```

Question 10

```
##add newly created variables to data frame
students.df<-data.frame(students.df,PartyAnimal,GPA.cat)

##export data as .csv
write.csv(students.df, file="new_students.csv", row.names = TRUE)
```

Question 11

```
fun.times<-students.df[which
                        (students.df$GPA<3.0 &
                            students.df$PartyNum>8 &
                            students.df$StudyHrs<15),]
nrow(fun.times)
```

```
## [1] 29
```

We have 29 students who have GPA less than 3.0, party more than 8 days a month, and study less than 15 hours a week.