Stat 6021: Guided Question Set 1

Download the dataset "students.txt" from Collab. The dataset contains information on students taking an introductory statistics class at a large public university in the early 2000s. The columns of the 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 2 alcoholic drinks
- `StudyHrs`: number of hours spent studying per week

For the questions below, you may use either the traditional approach or the dplyr approach (or even a combination of both approaches).

1. Looking at the variables above, is there a variable that will definitely not be part of any meaningful analysis? If yes, which one, and remove this variable from your data frame.
2. How many students are there in this data set?
3. How many students have a missing entry in at least one of the columns?
4. Report the median values of the numeric variables.
5. Report the mean and standard deviation of `StudyHrs` for female and male students.

6. Construct a 95% confidence interval for the mean `StudyHrs` for female students, and another 95% confidence interval for the mean `StudyHrs` for male students. Based on this intervals, do we have evidence that the mean `StudyHrs` is different between female and male students? **Hint:** use the `table()` function (base R) or the `count()` from the `dplyr` package to obtain the sample sizes of female and male students.

7. Compare the median `StudyHrs` across genders and `Smoke`.

8. Create a new variable called `PartyAnimal`, which takes on the value "yes" if `PartyNum` the student parties a lot (more than 8 days a month), and "no" otherwise.

9. Create a new variable called `GPA.cat`, which takes on the following values

   - "low" if GPA is less than 3.0
   - "moderate" if GPA is less than 3.5 and at least 3.0
   - "high" if GPA is at least 3.5

10. Add the variables `PartyAnimal` and `GPA.cat` to the data frame from part 1, and export it as a .csv file. Name the file `new_students.csv`. We will be using this data file for the next module.

11. Suppose we want to focus on students who have low GPAs (below 3.0), party a lot (more than 8 days a month), and study little (less than 15 hours a week). Create a data frame that contains these students. How many such students are there?