Inclass-activity

Q1

Q: 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.

Student Variable is the ID of students, which shouldnt be that important

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/Module1")

data = read.table("students.txt", header = TRUE)
data$Student <- NULL
head(data)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1 female    No      Yes      Yes 3.40        4        6        7
## 2 female    No       No       No 3.45        4        0       20
## 3   male    No       No      Yes 3.89        9        4       30
## 4 female    No       No       No 3.75        6        3       12
## 5   male   Yes      Yes      Yes 2.30       10       15       14
## 6 female   Yes      Yes       No 2.80        2        5       10
```

Q2

Q: How many students are there in this data set?

```
nrow(data)
```

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

Q3

How many students have a missing entry in at least one of the columns?

```
no_missing = data[complete.cases(data),]
nrow(data) - nrow(no_missing)
```

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

Q4

Report the median values of the numeric variables.

```
numeric_cols <- unlist(lapply(data, is.numeric))
numeric = data[, numeric_cols]
apply(numeric,2,median,na.rm=T)
```

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

Q5

Report the mean and standard deviation of StudyHrs for female and male students.

```
study_mean = mean(data$StudyHrs, na.rm = T)
study_std = sd(data$StudyHrs, na.rm = T)
print(c(study_mean, study_std))
```

```
## [1] 15.112450  9.490414
```

Q6

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.

```
female = data[data$Gender == "female",]
male = data[data$Gender == "male",]

confidence_interval <- function(data) {
  mean = mean(data$StudyHrs)
  num = nrow(data)
  sd = sd(data$StudyHrs)
  se = sd / sqrt(num)

  alpha = (1 - 0.95)
  degree.freedom = num - 1

  t.score = qt(p=alpha/2, df=degree.freedom,lower.tail=F)
  margin = t.score * se

  lower.bound <- mean - margin
  upper.bound <- mean + margin
  return(c(lower.bound,upper.bound))
}
male_ci = confidence_interval(male)
female_ci = confidence_interval(female)
print(c(male_ci,female_ci))
```

```
## [1] 12.71850 16.68535 13.93409 16.87970
```

Q7

Compare the median StudyHrs across genders and Smoke

```
female = data[data$Gender == "female",]
male = data[data$Gender == "male",]

f_smoke = female[female$Smoke == "Yes",]
f_no_smoke = female[female$Smoke == "No",]

m_smoke = male[male$Smoke == "Yes",]
m_no_smoke = male[male$Smoke == "No",]

median(female$StudyHrs)
```

```
## [1] 14
```

```
median(male$StudyHrs)
```

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

```
median(f_smoke$StudyHrs)
```

```
## [1] 10
```

```
median(f_no_smoke$StudyHrs)
```

```
## [1] 15
```

```
median(m_smoke$StudyHrs)
```

```
## [1] 14
```

```
median(m_no_smoke$StudyHrs)
```

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

Q8

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.

```
newData = data
newData$PartyAnimal <- ifelse(data$PartyNum>8, "yes", "no")
head(newData)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs PartyAnimal
## 1 female    No      Yes      Yes 3.40        4        6        7          no
## 2 female    No       No       No 3.45        4        0       20          no
## 3   male    No       No      Yes 3.89        9        4       30         yes
## 4 female    No       No       No 3.75        6        3       12          no
## 5   male   Yes      Yes      Yes 2.30       10       15       14         yes
## 6 female   Yes      Yes       No 2.80        2        5       10          no
```

Q9

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

```
newData$GPA.cat <- cut(data$GPA, breaks = c(0, 3, 3.5, 5),
labels = c("Low", "moderate", "high"), right=FALSE)
head(newData)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs PartyAnimal
## 1 female    No      Yes      Yes 3.40        4        6        7          no
## 2 female    No       No       No 3.45        4        0       20          no
## 3   male    No       No      Yes 3.89        9        4       30         yes
## 4 female    No       No       No 3.75        6        3       12          no
## 5   male   Yes      Yes      Yes 2.30       10       15       14         yes
## 6 female   Yes      Yes       No 2.80        2        5       10          no
##    GPA.cat
## 1 moderate
## 2 moderate
## 3     high
## 4     high
```

```
## 5      Low
## 6      Low
```

Q10

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.

```
result = read.table("students.txt", header = TRUE)
result$Student <- NULL
result$GPA.cat <- cut(data$GPA, breaks = c(0, 3, 3.5, 5),
labels = c("Low", "moderate", "high"))
result$PartyAnimal <- ifelse(result$PartyNum>8, "yes", "no")
write.csv(result, file="new_students.csv", row.names = TRUE)
```

Q11

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?

```
study_data = read.table("students.txt", header = TRUE)
study_data = study_data[complete.cases(study_data),]
study_data = study_data[study_data$PartyNum > 8,]
study_data = study_data[study_data$GPA < 3.0,]
study_data = study_data[study_data$StudyHrs < 15,]
nrow(study_data)
```

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