Alexander Rasho

1. Consider the data collected from the survey conducted during the first class in the attached .csv file. Perform the following steps. Give the corresponding R code (and answer other questions, show plots, if asked).

a. Load the survey data into a variable called “survey”

setwd("Downloads")

survey <- read.csv("survey_responses2.csv")

b. Which variables are numeric? Which variables are factors?

str(survey)

```
 - attr(*, "spec")=
  .. cols(
  ..   Timestamp = col_character(),
  ..   CS = col_double(),
  ..   Math = col_double(),
  ..   Statistics = col_double(),
  ..   ML = col_double(),
  ..   Domain = col_double(),
  ..   Communication = col_double(),
  ..   visualization = col_double(),
  ..   taken483 = col_character(),
  ..   plan483 = col_character(),
  ..   CSmajor = col_character(),
  ..   familiarR = col_double(),
  ..   familiarPython = col_double(),
  ..   hobbies = col_character()
  .. )
>
```

c. What is the mean value of Math skills in the class?

```
>attach(survey)
> mean(Math)
[1] 6.855072
```

d. Is the mean skill level in Math higher than that in CS?

```
> attach(survey)
> mean(CS)
[1] 7.173913
```

No, the mean skill level in Math lower than the mean skill level in CS.

e. How many students plan to take CPSC 483?

30 students will be taking CPSC 483 in the future.

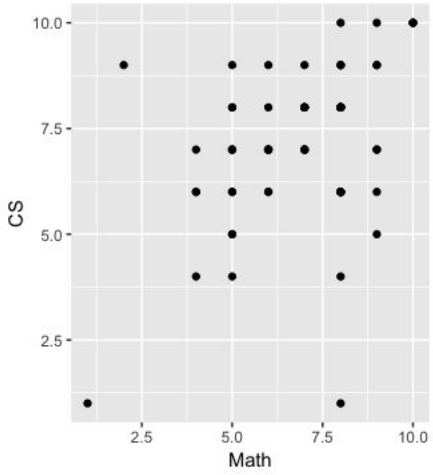
```
> attach(survey)
> table(plan483)
plan483
 No Yes
 39  30
```

f. Compare the levels of the factor variables, taken483 and plan483. Though both are answers to Yes/No questions, taken483 has a third level, the empty string (“”) because the question was not answered. Write R code to replace these empty string values with NA.

```
survey$taken483[survey$taken483==""] <-- NA
```

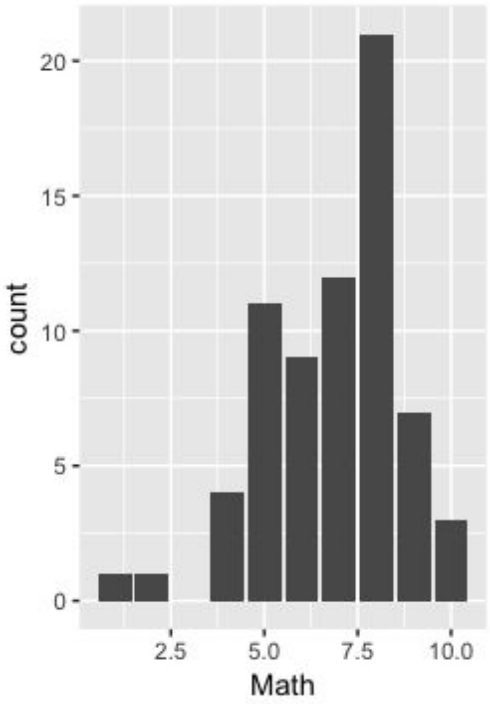
g. Plot (using ggplot) a scatterplot of variables “Math” and “CS”. Show both code and paste the plot as an image.

```
> attach(survey)
> library(ggplot2)
> ggplot(survey, aes(x=Math, y=CS)) + geom_point()
>
```



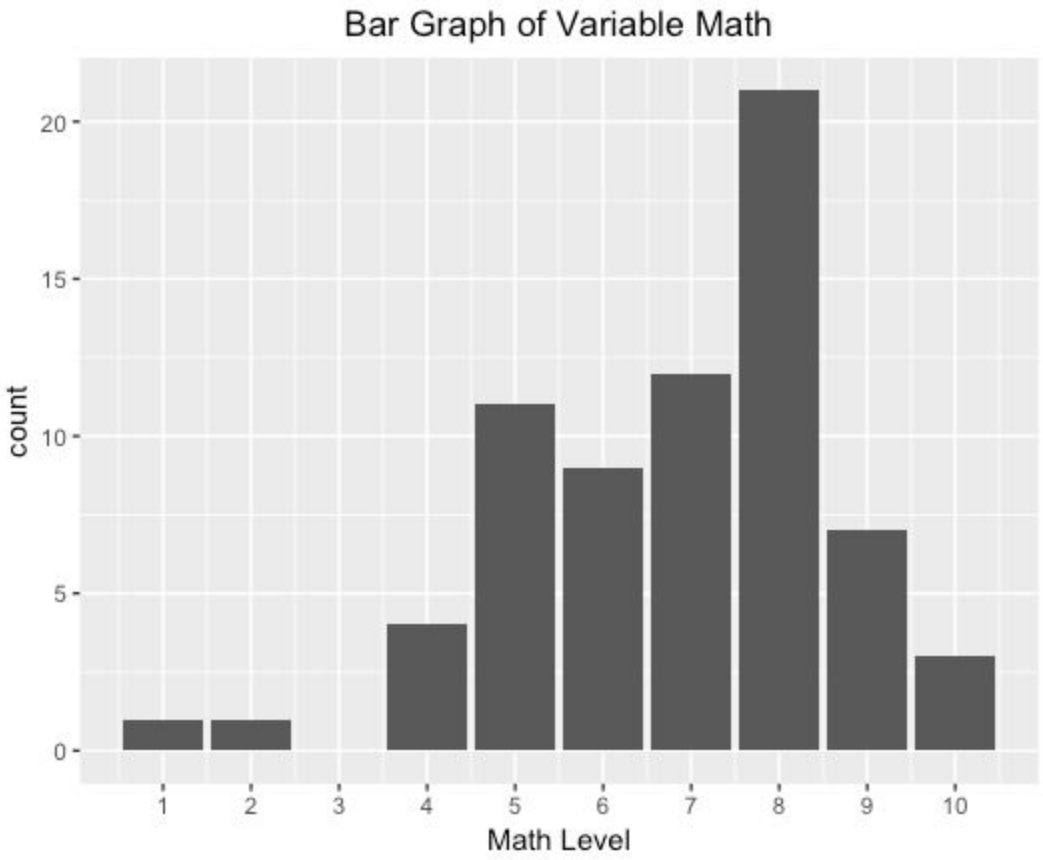
h. Plot (using ggplot) a bar graph of variable “Math”. Show both code and paste the plot as an image.

```
> attach(survey)
> library(ggplot2)
> ggplot(data=survey) + geom_bar(mapping = aes(x=Math))
```



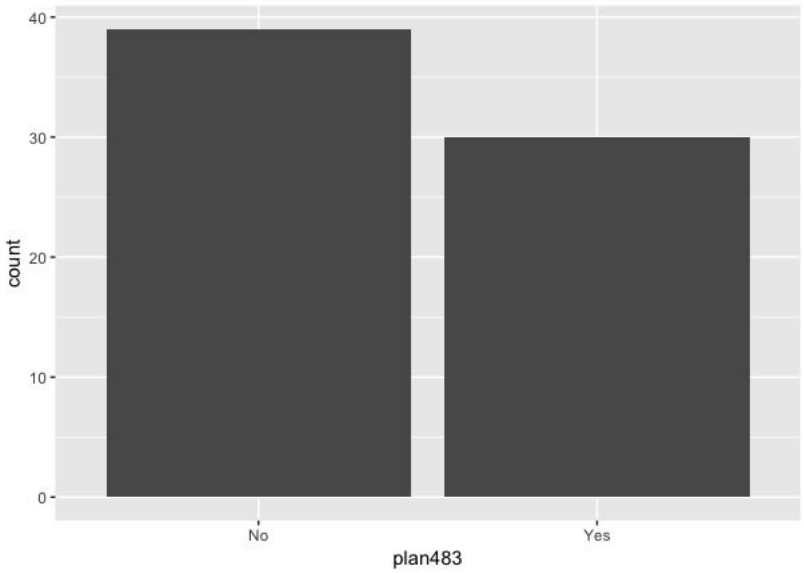
k. The plot above likely has x-axis labels not aligned with the bars. Provide your own breaks to match the variable values/bars. Also, add a plot title. Show both code and paste the plot as an image.

> ggplot(survey) + geom_bar(mapping = aes(x = Math)) + ggtitle("Bar Graph of Variable Math") + scale_x_continuous("Math Level", breaks = 0:10, labels = c("0","1","2","3","4","5","6","7","8","9","10")) + theme(plot.title = element_text(hjust = .5))



l. Plot (using ggplot) a bar graph of variable “plan483”. Show both code and paste the plot as an image.

```
> attach(survey)
> library(ggplot2)
> ggplot(data=survey) + geom_bar(mapping = aes(x=plan483))
```



2. Choose **one** of the following questions and answer in a short paragraph (5-6 sentences). Both questions will require some research on the Internet. Cite your sources.

a. (If you are familiar with Python:) What are the pros and cons for using R or Python as a programming language for data science?

Python is typically much faster to execute than R. R increases memory constraints due to being within a program itself. Python may be easier to learn since it's an easily (and sometimes already) understood language. Python is a multipurpose language that can easily integrate different components of code to have a smooth workflow. R seems like it's more visually appealing as well as having a community behind it that provides many packages to improve R.