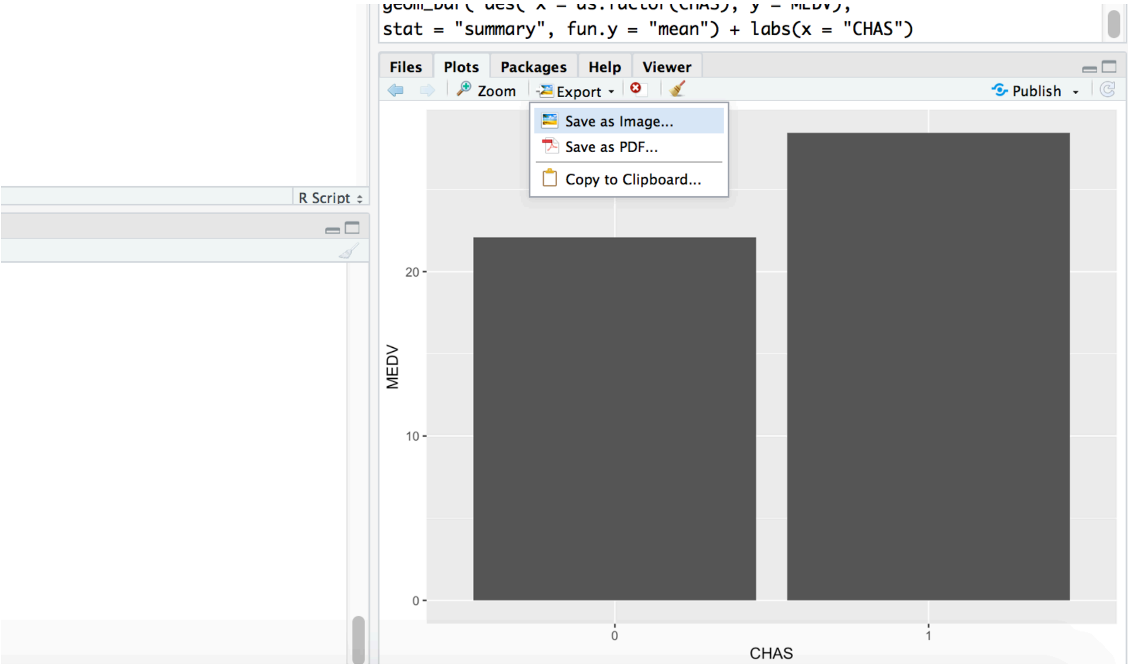
MSIS 510
Individual Assignment 1

Individual Assignment 1
Due at the beginning of the class on Nov. 7
Total: 8 points

The diamonds dataset that we will use in this application exercise consists of prices and quality information from about 54,000 diamonds, and is included in the ggplot2 package.

All the figures need to be generated using **ggplot2.** In RStudio, you can use **Export -> Sava as Image** as shown in the screenshot to save your plots.



```
geom_bar( aes( x = as.factor(CHAS), y = MEDV),
stat = "summary", fun.y = "mean") + labs(x = "CHAS")
```

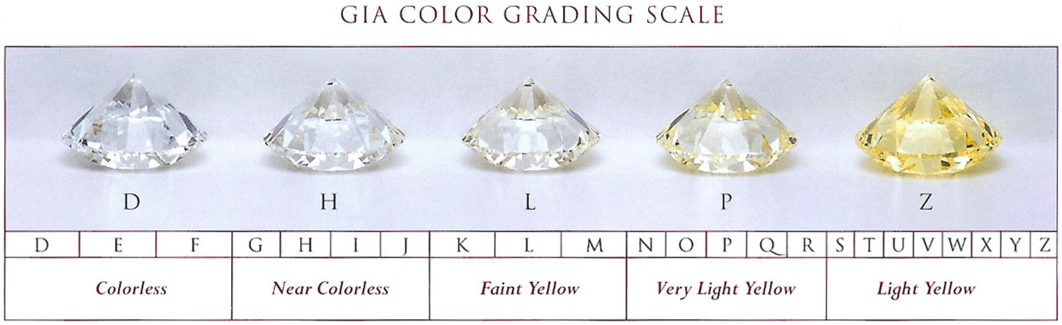
Paste your plots generated in RStudio to Word, answer all the questions in Word, and upload it to Canvas. You do not need to paste your R codes.

MSIS 510
Individual Assignment 1

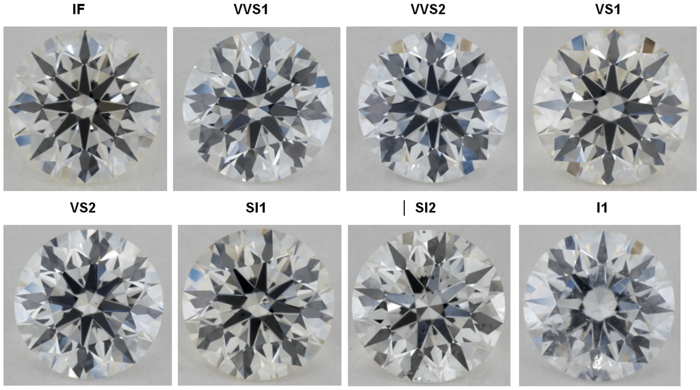
Since you already installed the ggplot2 in class, you don’t need to install them again. However each time you launch R you need to load the packages by *library(ggplot2).* To familiarize yourself with this dataset, you can use *?diamonds* to get more information and examine the dataset by *View(diamonds)*.

The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price (in 2008 $s): the four Cs (carat, cut, color, and clarity), as well as some physical measurements (depth, table, price, x, y, and z). Diamond carat weight is the measurement of how much a diamond weighs. A metric “carat” is defined as 200 milligrams. Cut grade is an objective measure of a diamond’s light performance, or, what we generally think of as sparkle.

The figure below shows color grading of diamonds:



The following figures illustrate the clarity grading scale:



MSIS 510
Individual Assignment 1

Questions:

1. Which variables are categorical? Which variables are numeric?

2. We would like to use data on the price and characteristics of many diamonds to figure out whether the price advertised for any given diamond is reasonable. With this regards, you need to create a scatterplot of price against carat weight. Put carat as the x-axis.

3. In the above figure, what do the data clustered in vertical lines in the plot tell us? Can you generate additional scatterplots by adding in a third dimension of information to justify your claims? Which variable (among cut, clarity, color) seems to better explain the additional variation of prices?

4. Create a bar chart of average prices across different cuts.

   There's something rather surprising about this plot - it appears that the lowest quality diamonds have the highest average price. We are going to figure out why in the next three questions.

5. Generate a bar chart, showing the total number of diamonds in the dataset, grouped by cut.

6. Generate a bar chart with cut as the x-axis, and illustrate the average carat weights under different cuts.

7. Based on the data exploration in Q5 and Q6 (and possibly one of the scatterplot you generated in Q3), can you give an explanation on the counterintuitive finding in Q4?