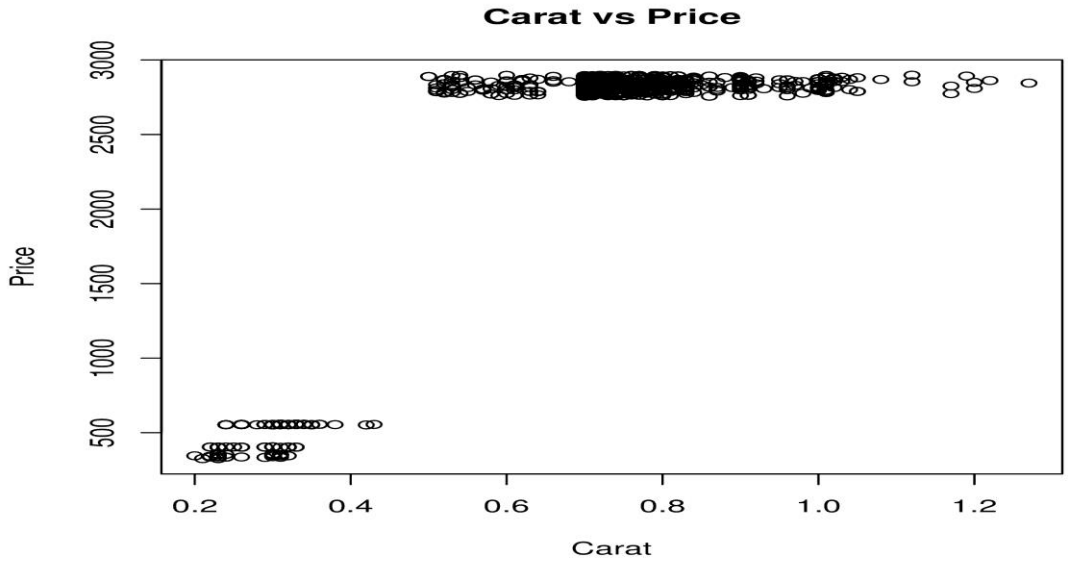
```
# HW 1 Due Monday Sept 18, 2017. Upload R file or notebook to Moodle with
# filename: HW1_490ID_41.R
# Do Not remove any of the comments. These are marked by #

###Your Unique Class ID: 41
# 1. Load the data for this assignment into your R session with the following command:

library(ggplot2)
data(diamonds)

#(1). Check to see that the data were loaded by running:
objects(diamonds)

# This should show ten variables:  carat, clarity, color, cut, depth,  price, table, x, y, z
#(2). Use the nrow() function to find out how many observations there are.
nrow(diamonds) #There are 53940 observations.

# For the following questions, use one of: head(), summary(),  class(), min(), max(), hist(),
# quantile(), table(), to answer the questions.
#(3). Show the summary or the structure of this dataset.
summary(diamonds)

#(4). List the categorical variables in this dataset.
my.cat.vars = diamonds[,c("cut","color","clarity")]
my.cat.vars
#After looking at summary, this displays the categorical variables cut , color and clarity.

#(5). What was the highest price of the diamonds ?
max(diamonds$price)  ###   The highest price is 18823.
#(6). What was the average price of the diamonds ?
mean(diamonds$price)  ###  The average price 3932.8

#(7). What is the number of the Ideal cut ?
summary(diamonds$cut)   ###  The  number of the Ideal cut is 21551

#(8). What is the number of the diamonds which are Premium and have a clarity level
#     of IF?
summary(diamonds$cut=="Premium" & diamonds$clarity=="IF") ### The number is  230.

#(9). What is the average price difference between the clarity level SI2 and IF?
#     Hint(Use aggregate())
aggregate(diamonds$price, by=list(diamonds$clarity),FUN=mean)
avg.prices$x[avg.prices$Group.1 == 'SI2'] - avg.prices$x[avg.prices$Group.1 == 'IF']
### The average price difference 2198.189

#(10). Total depth percentage is represented as the depth divided by
```

```
#      the mean of the length and width of the diamond,
# (11).
# Try running each expression in R.
# Record the error message in a comment
# Explain what it means.
# Be sure to directly relate the wording of the error message with the problem you find in the
expression.
z/mean(x, y)
### Error: object 'z' not found
### This error appears since the object z is not loaded by itself but part of the data.frame
### To access it, we must use diamonds$z.

diamonds$z/(mean(diamonds$x, diamonds$y))
### Error in mean.default(diamonds$x, diamonds$y) : 'trim' must be numeric of length one
### Since the second option for the function mean() is trim it must be a fraction not a vector.

diamonds$z/(rowMeans(diamonds$x, diamonds$y))
### Error in rowMeans(diamonds$x, diamonds$y) : 'x' must be an array of at least two
dimensions.
### When you look at the documentation, the first input of rowMeans() ustve an array of two or
more dimensions. So, we must concatenate both vectors first and use that as our input.

#(12).Study the following code about how to do the compution
# calculation that we want for the previous question.
Depth <- (diamonds$z)/rowMeans(cbind(diamonds$x, diamonds$y))

#(13). Can you get the same result without using function?
#              Given an expression to it.  Name the values Depth1
my.avg = (diamonds$x+ diamonds$y)/2
Depth1 = diamonds$z/my.avg


#(14) What did you get from all.equal(Depth, Depth1)
###  TRUE, thus we can get the same result without using the function.

# 2. Run the following code to make a plot.
# (don't worry right now about what this code is doing)
plot(diamonds[1:1000,]$carat, diamonds[1:1000,]$price, xlab = "Carat", ylab = "Price", main =
"Carat vs Price")

# (1) Use the Zoom button in the Plots window to enlarge the plot.
# Resize the plot so that it is long and short, so it is easier to read.
# Include this plot in the homework your turn in.
```



(2) Make an interesting observation about the relation between
Carat and Price based on this plot
(something that you couldn't see with the calculations so far.)
Looking at the plot, I see that after approxiately 0.5 carats the price skyrockets and mostly fall within
the range of 2,500 and 3,000. So, if I ever get married I should get a higher carat ring that falls in the lower range of that price interval.

(3) What interesting question about the diamonds
would you like to answer with these data, but don't yet know
how to do it?
I would like to if there is a way to determine which factor (carat, cut, clarity or depth)or combination of factors influences price the most.

For the remainder of this assignment we will work with
one of the random number generators in R.

4.
Use the following information about you to generate some random values:
#a. Use you UIN number to set the seed in set.seed() function.
set.seed(645914315)

#b. Use your birthday month for the mean of the normal.
my.mean = 10

#c. Use your birthday day for the standard deviation (sd) of the normal curve.
my.sd = 19

#d. Generate 10 random values using the parameters from b and c.
rnorm(10,mean = my.mean, sd = my.sd)

#e. Assign the values to a variable named with your first name.
Vanessa = rnorm(10,my.mean, my.sd)
[1] -1.811788 -20.001788 31.201479   6.251553 -9.665065 -6.228426 28.452773
[8] 12.579398   5.775780 15.493559

#f. Provide/Show the values generated.
Vanessa
[1] 32.180251 10.325125 -14.988669 21.420551 -5.353224 13.378861   8.080348
[8] 15.782661 44.379324   4.345210

5.
#(1). Generate a vector called "normsamps" containing
# 100 random samples from a normal distribution with
# mean 5 and SD 2.
normsamps = rnorm(100, mean=5,sd=2)

#(2). Calculate the mean and sd of the 100 values.
mean(normsamps)
sd(normsamps)
The mean of norsmamps is 4.993774 and standard deviation was 1.975665

(3). Use implicit coercion of logical to numeric to calculate
the fraction of the values in normsamps that are more than 8.
length(normsamps[normsamps > 8])/length(normsamps) #This gives 0.04

(4). Look up the help for rnorm.
? rnorm

You will see a few other functions listed.
Use one of them to figure out about what answer you
should expect for the previous problem.
That is, find the area under the normal(5, 2) curve
to the right of 8. This should be the chance of getting
a random value more than 8.

# What value do you expect?
pnorm(8, mean = 5, sd =2, lower.tail =FALSE)
pnorm(), computes the probability X >8. So, I expect a value of 0.0668072.

What value did you get?
I got in number (3) the value of 0.04.

Why might they be different?
We are using a 100 samples which is fairly small because our answers are just approximations we can expect them to be different.