STA303/1002H1S: Methods of Data Analysis II **Winter 2017**

Assignment 0 : Visualizing Data Using ggplot2

Out: Jan. 05, 2017 *Due: Jan. 21, 2017*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 10 :00pm, Jan. 21, 2016. Submit your solution as instructed by Crowdmark, namely, one pdf file for each question.

Late assignments will be subject to a deduction of 5% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. A Rmarkdown template for the solution is provided. Please following the instruction in that template to complete your solution. You should produce a PDF file, split the PDF into different files to get your solution PDF for each question. Also, make sure the source R code at the end is complete. Marks will be deducted if the instructions herein are not followed.

Data

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows : A data frame with 53,940 rows and 10 variables :

The variables in the dataset are :

- price : price in US dollars ($326-$18,823)
- carat : weight of the diamond (0.2-5.01)
- cut : quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color : diamond colour, from J (worst) to D (best)
- clarity : a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
- x : length in mm (0-10.74)
- y : width in mm (0-58.9)
- z : depth in mm (0-31.8)

To access this data set, you could install R package *ggplot2* using the following R command

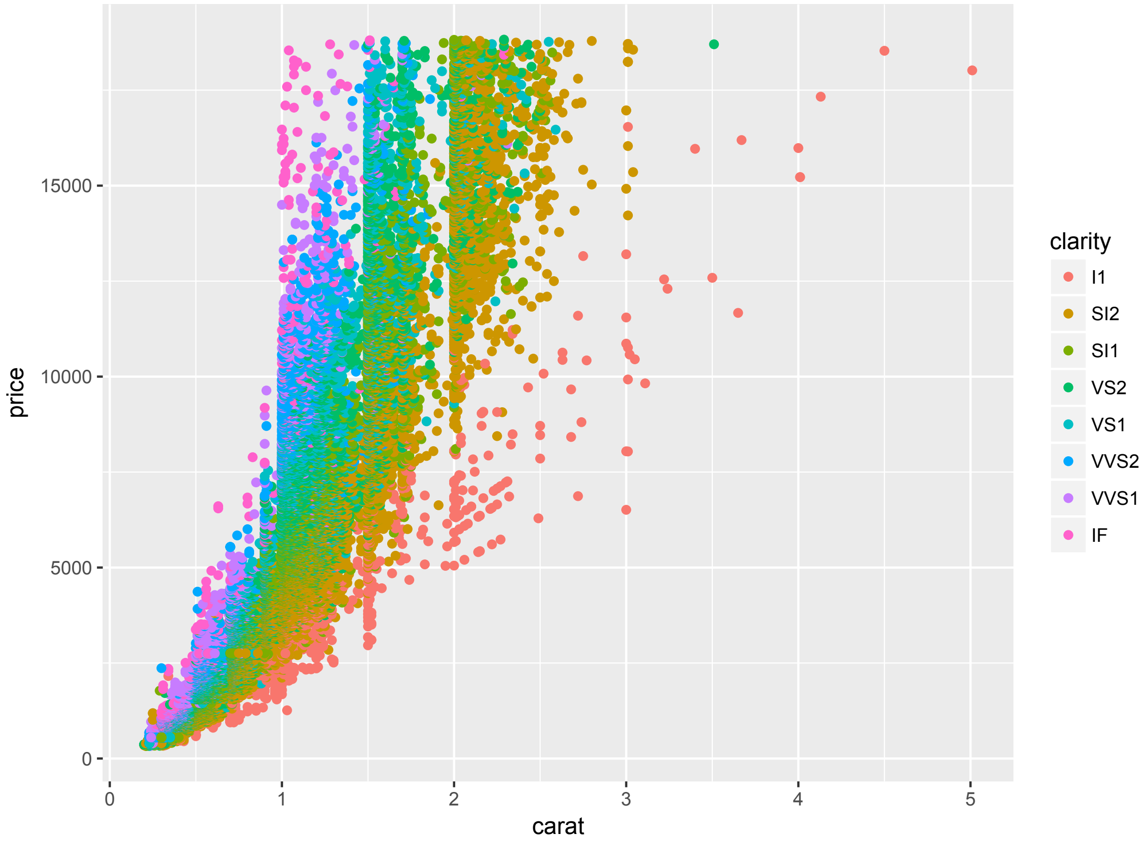
```r
# Install ggplot2 R packages
install.packages("ggplot2")

# load in Diamonds data
library(ggplot2)
data(diamonds)

# see how many rows in the datasets
nrow(diamonds)
```

Listing 1 – Install package and read in data

Questions

Using R to do all the analysis for the following. Uncompleted R code is given at the end of the assignment.

1. (1 point) How many factor variables in this data set ? Use R command *str(diamonds)* to find it. For each factor variable, find the one-way frequency table for it. An example of **cut** variable is given in the solution template.
2. (1 point) Produce the following plot and show it in your solution. Give a brief comment for what you learn from this plot. In your Rmarkdown file, make sure your set the **echo=TRUE** in your R chunk, so the R code will be presented in your turned in solution too.



Does it look like a bunch of flowers ? Happy new year, students :)

To produce above plot, you could either use ggplot() or qplot() in the ggplot2 package. To learn more about ggplot2, two good introductory tutorials are :
— By Josef Fruehwald : `http://www.ling.upenn.edu/~joseff/avml2012/`
— Hadley's GitHub : `https://github.com/tidyverse/ggplot2/wiki`

Incomplete R code

```r
#
# R code for STA303 or STA1002H1s Assignment 0
# date: Jan. 02, 2017
#

## Load in the data set
library(ggplot2)
data(diamonds)

## a quick look of this data set
head(diamonds)
str(diamonds)
summary(diamonds)
dim(diamonds)

## Q1: Find number of factor variables in the data set
str(diamonds)

# percentage of each level of variable "cut" and round off to 4 digits
round( summary(diamonds$cut)/nrow(diamonds), digits=4)

# ... for other variables ...

## Q2: Give plot and write comment
# Consider using ggplot,
ggplot()

qplot()
```

Listing 2 – Incomplete code for the data analysis