STAT 217- PROJECT 4
Due Monday Nov 23rd beginning of class

The website `www.adiamondisforever.com` educates the layperson on the factors that influence the price of a diamond stone. These are the 4 C's: Carat, Clarity, Colour and Cut. In this project, you will focus on 2 of the 4 C's: Carat and Colour.

The weight of a diamond stone is indicated in terms of carat units. One carat is equivalent to 0.2 grams. All other things being equal, larger diamond stones command higher prices in view of their rarity.

The most prized diamonds display colour purity. They are not contaminated with yellow or brown tones. Top colour purity attracts a grade of D. Subsequent degrees of colour purity are rated E, F, G, F all the way down the alphabet ladder.

To assist shoppers, independent certification bodies assay diamond stones and provide each of them with a certificate listing their caratage and their grades of clarity, colour and cut. We will be using data presented in a newspaper advertisement. The newspaper advertisement however only provided, for each stone, details on the certification body and its assessment of the caratage, clarity, and colour of the stones. Three certification bodies were mentioned in the advertisement, namely New York based Gemmological Institute of America (GIA) and Antwerp based International Gemmological Institute (IGI) and Hoge Raad Voor Diamant (HRD). Their reputations could be a factor in the pricing of the diamond stones. The data were found at: `http://www.amstat.org/publications/jse/v9n2/4Cdata.txt` and are available for you on d2l.

We are interested in the relationship between caratage and price (denoted in Singapore dollars), but we suspect that this relationship changes depending on the color of the diamond. Write a report following the Project Writing Guidelines. Here are some additional guidelines you should follow.

Introduction
Your question of interest should address whether the slope adjustment is necessary for different colors or not.

Statistical Procedures Used

- Be sure to describe the patterns you see in the initial scatterplots
- Be sure to justify the log transformation
- Evaluate the model assumptions after the log transformation
  - When considering the assumption of linearity- remember, we want the relationship between log(carat) and log(price) to be linear for each color.
  - Why don't we have to worry about multicollinearity?

Summary of Statistical Findings: This section should have seven sentences.

- This first sentence is your evidence sentence and should directly answer your question of interest.
- The next six sentences will be estimate sentences for each of the six colors. Describe the estimated relationship between Carat and Price (on the original scale) for each of the six colors.

Further Analysis: This section is worth 10 points, don't forget to do it!
Draw, by hand, the estimated regression line for each of the six colors. Remember, the estimated regression line is on the log scale, so the axes of your plots will be labeled log(Carat) and log(Price). Write the color, slope, and y-intercept of the line in the upper right hand corner of each plot.

You should have six hand-drawn plots in total. Each plot should be labeled with a different color, slope, and y-intercept. You will not receive many points if you do not clearly define the color, slope, and y-intercept of your plots.

Additionally, answer the following question: What is the estimated price of a diamond that weighs 0.71 carats and has color G? Remember that your estimated regression line is on the log scale!

Scope of Inference: Some of you are still having trouble with this section. This will help you:

- Sentence one: Were diamonds in the study randomly selected from a larger population of diamonds?
- Sentence two: Were diamonds in the study randomly assigned to carats and colors?

```
##You only need to run this first chunk of code once after importing the data##
diamonds <- diamonds[,c(2,3,6)]
diamonds$Color <- as.factor(diamonds$Color)
require(mosaic)
##You may need to install the lattice package##
require(lattice)
options(show.signif.stars = F)


#Look at and describe scatterplots
pairs(diamonds)
xyplot(Price~Carat|Color,data=diamonds)
xyplot(log(Price)~log(Carat)|Color,data=diamonds)

#Why should you consider a log transformation?
fit.first <- lm(Price~Carat*Color, data = diamonds)
plot(fit.first, which = 1, pch = 20)

#Look at the interaction model
fit.Int <- lm(log(Price)~I(log(Carat))*Color, data = diamonds)
summary(fit.Int)
anova(fit.Int)
par(mfrow=c(2,2))
plot(fit.Int)
xyplot(log(Price) ~ log(Carat) | Color, type = c("p","r"),data=diamonds)

fit.int.cellmeans <- lm(log(Price)~Color+I(log(Carat)):Color-1, data = diamonds)
confint(fit.int.cellmeans)
```