STAT 217- PROJECT 3
Due Tuesday, July 29 at the beginning of class

Write up the results using a word processor and include appropriate plots and output in your report. Part of the grade is based on the write-up, so you do need to explain your results as if you are talking to a classmate who knows some statistics but does not know anything about computer output.

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.

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.

Being products of Nature, diamonds have birthmarks or inclusions only visible under a jeweller's magnifying glass or a microscope. Diamonds with no inclusion under a loupe with a 10 power magnification are labelled IF (internally flawless). Lesser diamonds are categorised in descending order as very very slightly imperfect (VVS1 or VVS2) and very slightly imperfect (VS1 or VS2).

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.

The cut (or faceting) of a raw diamond stone relies on the experience and the craftsmanship of the diamond cutter. The optimal cut should neither be too deep nor too shallow for it will impede the trajectory of light and thereby the brilliance or fire of a diamond stone.

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 newaspaper 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.

Given the information in the dataset, a multiple linear regression (MLR) model is a natural path to explore. Generally speaking, one would expect the price (denoted in Singapore dollars) of a stone to move in tandem with the caratage. However, the relationship may not be linear as heavier stones are more prized than the lighter ones. The data were found at: `http://www.amstat.org/publications/jse/v9n2/4Cdata.txt` and are avaialbe for you on d2l.

- Look at a scatterplot with `carat` as the explanatory variable and `price` as the response variable. Also look at the residual vs. fitted values plot to assess the 'straight enough condition' and the constant variance assumption for a SLR. Do these data require a transformation?

```
require(mosaic)
options(show.signif.stars = F)
#read the data into R and name it diamonds
scatterplot(Price~Carat, data = diamonds, boxplots = F, smooth = F, pch = 20, col = "black")
fit.SLR <- lm(Price~Carat, data = diamonds)
plot(fit.SLR, which = 1, pch = 20)
```

- Provide a scatter plot of `Carat` on the x-axis and `log(Price)` on the y-axis and code the points based on `Color`.

```
scatterplot(log(Price) ~ Carat|Color,diamonds, smooth = F, boxplots = F, reg.line = F)
```

- Fit an interaction model with `Carat`, and `Color` as explanatory variables and `log(Price)` as the response. Write out the estimated SLR equation for each Color.
- Interpret the regression coefficients (intercept and slope) in the context of the problem and on the original scale for diamonds with color G.
- Do you think the slope adjustment is necessary for different colors or not? Justify your answer. Look at the appropriate test from an ANOVA of the interaction model fit and interpret the results. *Make sure you state the hypotheses you are testing, the F-statistic, it's distribution under H_0, the p-value and your decision in the context of the problem.* If you decide the interaction terms are unnecessary, fit the additive model and proceed with LRT until you choose a 'best' model.
- Visualize the fit of your model by adding the SLR lines for each single color to the scatterplot of the data.

```
scatterplot(log(Price) ~ Carat|Color,diamonds, smooth = F, boxplots = F, lwd = 2)
```

- **Extra Credit:** Look at the diagnostic plots and assess the assumptions and conditions for running a MLR. Why didn't we have to worry about multicollinearity?

R CODE

```
require(mosaic)
options(show.signif.stars = F)
#read the data into R and name it diamonds
scatterplot(Price~Carat, data = diamonds, boxplots = F, smooth = F, pch = 20, col = "black")
fit.SLR <- lm(Price~Carat, data = diamonds)
plot(fit.SLR, which = 1, pch = 20)
scatterplot(log(Price) ~ Carat|Color,diamonds, smooth = F, boxplots = F, reg.line = F)
fit.Int <- lm(log(Price)~Carat*Color, data = diamonds)
summary(fit.Int)
anova(fit.Int)
fit.Add <- lm(log(Price)~Carat+Color, data = diamonds)
summary(fit.Add)
Anova(fit.Add)
```