Luke Beebe
STAT 463 – Assignment 5 [code is in R file]

1) To start the project, I loaded diamonds_data.txt into R and fit a linear model to it where the response variable is the diamond price, and the explanatory variable is the clarity rating. The model equation for diamond price that I got is:

   price=2694.8+2362.3(VS1)+3163.4(VS2)+2872.9(VVS1)+2661.8(VVS2)

2) The estimate of the mean price of a diamond within each clarity rating level is listed below:

   IF=2694.8, VS1=5057.1, VS2=5858.2, VVS1=5567.7, VVS2=5356.6

   By a glance, it seems IF's mean price vastly differs from the other clarity ratings. I wonder if there's a statistical backing.

3) The method to use to determine whether there is a statistically significant relationship between diamond price and clarity rating is to check the p-value of the model, which is extremely low at 2.216e-05. This means there is a statistically significant relationship between the variables. However, one thing to consider is how inaccurate it is at prediction. The R^2 value is 0.08, which is very low, meaning it's inaccurate at predictions.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2694.8      494.2   5.453 1.03e-07 ***
CLARITYVS1    2362.3      613.9   3.848 0.000145 ***
CLARITYVS2    3163.4      668.5   4.732 3.42e-06 ***
CLARITYVVS1   2872.9      671.4   4.279 2.52e-05 ***
CLARITYVVS2   2661.8      618.0   4.307 2.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3278 on 303 degrees of freedom
Multiple R-squared:  0.08428,   Adjusted R-squared:  0.0722
F-statistic: 6.972 on 4 and 303 DF,  p-value: 2.216e-05
```

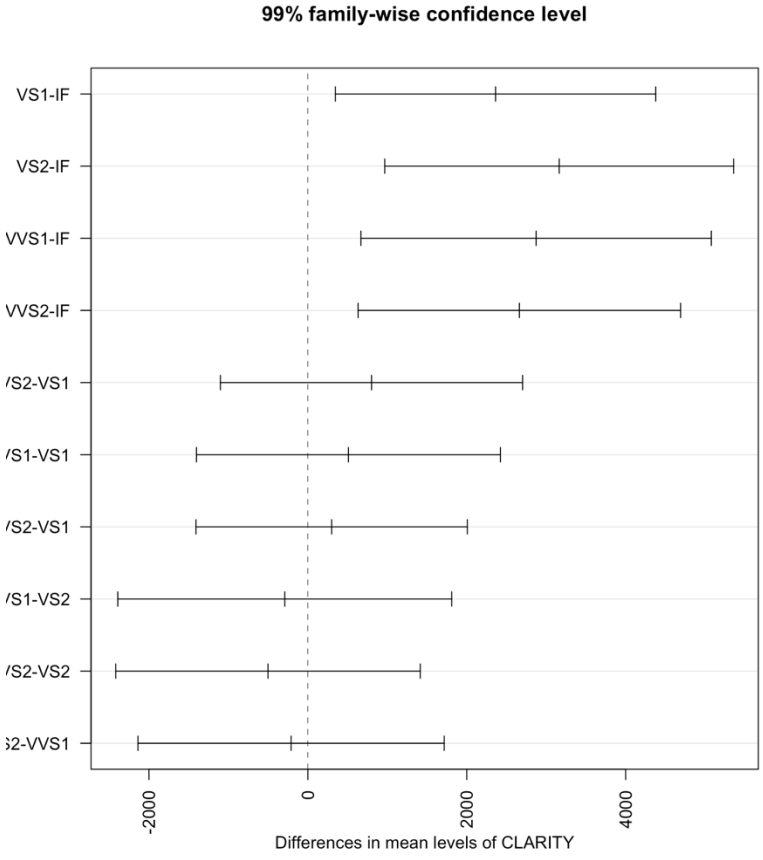
4) To compare which clarity rating's mean price is statistically different from the others I deployed multiple methods. The family-wise error rate (probability of getting a type 1 error on one hypothesis test) is calculated as $1-(1-a)^n$ where a=alpha and n=num of groups comparing. If we'd like to keep the error rate below 0.1, then an a<=0.02. The other method I used was bonerroni's correction method, which takes the regular alpha, a=0.05, and divides by n, so a=0.01. Luckily, question 5 asks for a 99% confidence interval, so I used a=0.01 to be safe while comparing the means between groups. At a=0.01, the family-wise error rate is approximately 0.049. I also calculated a pairwise t-test between the groups out of curiosity to see if it would give us the same results. The p-values were slightly different.

```
> data.tukey
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = PRICE ~ CLARITY)

$CLARITY
               diff        lwr      upr     p adj
VS1-IF    2362.2643   346.2595 4378.269 0.0013599
VS2-IF    3163.3971   967.9297 5358.864 0.0000335
VVS1-IF   2872.8619   667.8395 5077.884 0.0002434
VVS2-IF   2661.7786   632.1729 4691.384 0.0002162
VS2-VS1    801.1328 -1100.7220 2702.988 0.6388421
VVS1-VS1   510.5976 -1402.2794 2423.475 0.9053493
VVS2-VS1   299.5142 -1408.1959 2007.224 0.9785012
VVS1-VS2  -290.5352 -2391.7016 1810.631 0.9911977
VVS2-VS2  -501.6185 -2417.8846 1414.648 0.9113017
VVS2-VVS1 -211.0833 -2138.2892 1716.122 0.9964082
```

```
> pairwise.t.test(PRICE,CLARITY,p.adjust.method="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  PRICE and CLARITY

     IF      VS1     VS2     VVS1
VS1  0.00145 -       -       -
VS2  3.4e-05 1.00000 -       -
VVS1 0.00025 1.00000 1.00000 -
VVS2 0.00022 1.00000 1.00000 1.00000
```

Both models show that the means between IF (internally flawless) and all other groups are statistically significant at both a=0.01 and a=0.02.

5) Below I constructed two-sided 99% confidence intervals between each group. The groups that contain 0 are the groups whose means aren't statistically different from part 4.



Bonus:

I found it peculiar that the 'nicest' (internally flawless) diamonds were the cheapest. I did some digging and found that they on average weighed the least of all the clarities.