Example4_5

Kevin Cummiskey

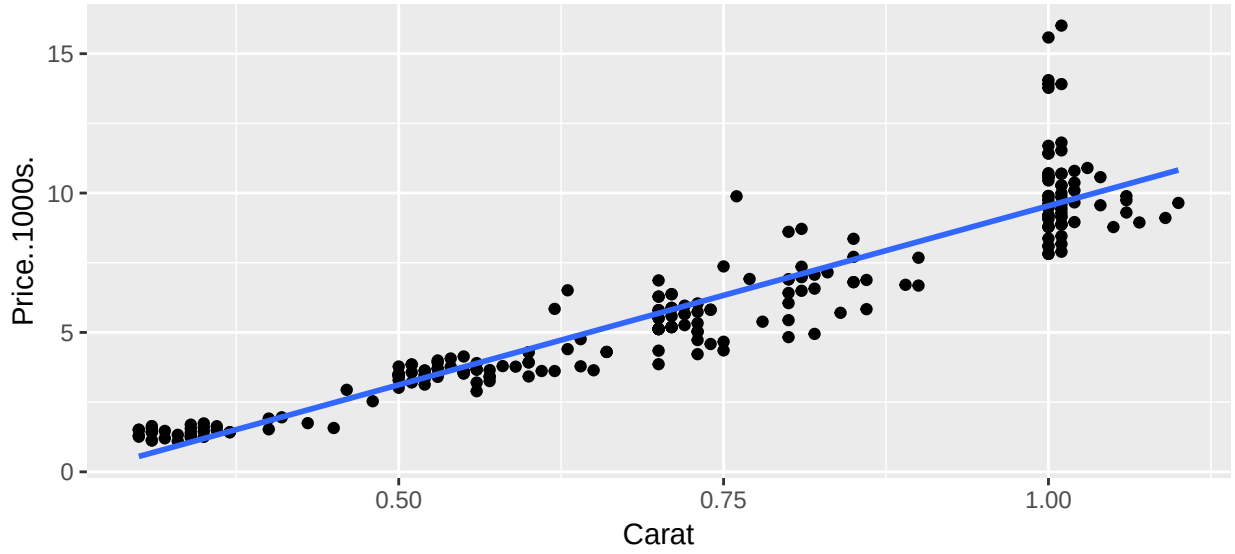
October 28, 2019

```
diamonds = read.table(file = "http://www.isi-stats.com/isi2/data/diamonds.txt",
                      header = T)
```

One variable analyses

Price vs Weight

```
diamonds %>% ggplot(aes(x = Carat, y = Price..1000s.)) +
  geom_point() + geom_smooth(method = "lm", se = F)
```




```
model_weight = lm(Price..1000s. ~ Carat, data = diamonds)
summary(model_weight)
```

```
##
## Call:
## lm(formula = Price..1000s. ~ Carat, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2819 -0.6242 -0.0978  0.3977  6.3380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.3010     0.2543  -12.98   <2e-16 ***
## Carat        12.8426     0.3355   38.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.195 on 228 degrees of freedom
## Multiple R-squared:  0.8653, Adjusted R-squared:  0.8647
## F-statistic:  1465 on 1 and 228 DF,  p-value: < 2.2e-16
```

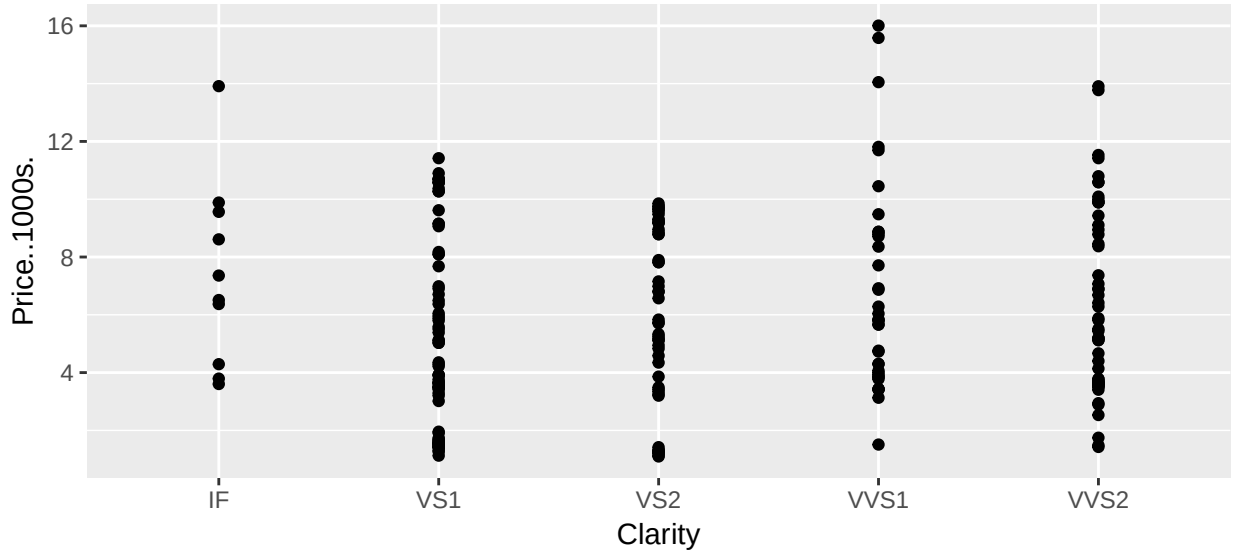
```
anova(model_weight)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Carat       1 2092.26 2092.26  1465.1 < 2.2e-16 ***
## Residuals 228  325.59    1.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What would you conclude from this model?

Price vs clarity

```
diamonds %>% ggplot(aes(x = Clarity, y = Price..1000s.)) +
  geom_point()
```



How many indicator variables do we need?

Here is the model:

$$y_i = \beta_0 + \beta_1 VS1_i + \beta_2 VS2_i + \beta_3 VVS1 + \beta_4 VVS2 + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

What is the reference category?

```r
model_Clarity = lm(Price..1000s. ~ Clarity, data = diamonds)
summary(model_Clarity)
```

```
##
## Call:
## lm(formula = Price..1000s. ~ Clarity, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2052 -2.6522 -0.7788  2.5254  9.2928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.390      1.016   7.278 5.62e-12 ***
## ClarityVS1    -2.246      1.082  -2.076   0.0391 *
## ClarityVS2    -1.489      1.111  -1.341   0.1814
## ClarityVVS1   -0.675      1.141  -0.591   0.5548
## ClarityVVS2   -1.102      1.101  -1.001   0.3179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.211 on 225 degrees of freedom
## Multiple R-squared:  0.04039,    Adjusted R-squared:  0.02333
## F-statistic: 2.368 on 4 and 225 DF,  p-value: 0.05363
```

```r
anova(model_Clarity)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Clarity     4   97.66  24.415  2.3676 0.05363 .
## Residuals 225 2320.19  10.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What would you conclude from this model?