One way to analyze how the variables carat, color, cut, and clarity are related is to check for multicollinearity. There are two different methods of determining if there is multicollinearity between these predictor values. The first method of checking for multicollinearity is to review the summary of the linear regression with all the predictors included. If the model has a significant ANOVA F test and insignificant t tests, that is one sign for multicollinearity. Below are the summary results after fitting the model with the predictor values of carat, color, cut, and clarity and the response variable of price.

```
Call:
lm(formula = price ~ carat + color + cut + clarity)

Residuals:
    Min      1Q  Median      3Q     Max
 -95976   -4234    1186    4379 1932205

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9961.34     538.52  18.497  < 2e-16 ***
carat         24512.88      45.76 535.640  < 2e-16 ***
colorE         -240.19     112.43  -2.136   0.0326 *
colorF         -846.05     111.88  -7.562 3.97e-14 ***
colorG        -2117.49     113.88 -18.594  < 2e-16 ***
colorH        -3109.97     120.80 -25.745  < 2e-16 ***
colorI        -3328.87     122.70 -27.131  < 2e-16 ***
colorJ        -5311.58     142.94 -37.159  < 2e-16 ***
cutGood       -1866.38     283.59  -6.581 4.68e-11 ***
cutIdeal       1634.42     258.44   6.324 2.55e-10 ***
cutVery Good  -1333.65     261.94  -5.091 3.56e-07 ***
clarityIF    -18567.14     496.50 -37.396  < 2e-16 ***
claritySI1   -23210.90     479.61 -48.395  < 2e-16 ***
claritySI2   -23752.02     483.99 -49.076  < 2e-16 ***
clarityVS1   -21794.88     480.01 -45.405  < 2e-16 ***
clarityVS2   -22571.31     480.41 -46.984  < 2e-16 ***
clarityVVS1  -20138.46     482.18 -41.766  < 2e-16 ***
clarityVVS2  -20688.67     481.81 -42.939  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14790 on 210620 degrees of freedom
Multiple R-squared:  0.5829,    Adjusted R-squared:  0.5829
F-statistic: 1.732e+04 on 17 and 210620 DF,  p-value: < 2.2e-16
```

Based on this output, it appears both the ANOVA F test and the t tests are very significant. This is one sign that there is not multicollinearity between the predictors. This also shows the model is very fitted, but it is possible it is overfitted and not all the predictors are required.

The next method of checking for multicollinearity is by calculating the variance inflation factor (VIF) for each of the predictors. The goal is for the predictors to have a VIF below 10. If the VIF exceeds 10, then it is a sign there is multicollinearity. Below are the VIF for the predictors in the full model.

```
     carat       colorE       colorF       colorG       colorH       colorI       colorJ      cutGood     cutIdeal cutVery Good
  1.032792     1.724729     1.744493     1.705172     1.591938     1.569709     1.374450     4.935294    15.286019    13.991103
clarityIF   claritySI1   claritySI2  clarityVS1   clarityVS2  clarityVVS1  clarityVVS2
10.468867    35.654328    23.445641    32.680147    32.001801    26.207331    26.998368
```

According to this calculation, there is significant collinearity between the different types of clarity, and there appears to be some collinearity within the Ideal and Very Good cuts. As noted in the exploratory data section, both cut and clarity have one high quality and very rare category: Astor Ideal and Flawless, respectively. However, these rare categories are not showing up as they are currently set as the reference category. They also have significantly fewer values in those categories, so it is important to make sure the VIFs are not high only because the reference category is very small. I will change the reference category to Good for cut and SI1 for clarity and re-calculate the VIFs. Below are the values with the adjusted reference categories.

```
       carat        colorE        colorF        colorG        colorH        colorI        colorJ cutAstor Ideal       cutIdeal
    1.032792      1.724729      1.744493      1.705172      1.591938      1.569709      1.374450      1.220156      3.903726
cutVery Good    clarityFL    clarityIF   claritySI1   clarityVS1   clarityVS2  clarityVVS1  clarityVVS2
    3.797570      1.055370      1.337017      2.173201      2.083740      2.058656      1.869693      1.893783
```

With the new reference categories, the VIFs have reduced and none exceed 10. The VIFs were high due to the low proportion of diamonds in the previous reference categories. Therefore, it is appropriate to conclude there is not multicollinearity between carat, cut, color, and clarity.